
Article:
Marsden, Emma Josephine orcid.org/0000-0003-4086-5765, Morgan-Short, Kara, Thompson, Sophie et al. (1 more author) (2018) Replication in second language research : Narrative and systematic reviews, and recommendations for the field. Language Learning. pp. 321-391. ISSN 0023-8333

https://doi.org/10.1111/lang.12286

LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

SYSTEMATIC REVIEW ARTICLE

Replication in Second Language Research: Narrative and Systematic Reviews and Recommendations for the Field

Emma Marsden [a], Kara Morgan-Short [b], Sophie Thompson [a], and David Abugaber [b]

[a]University of York and [b]University of Illinois at Chicago

A note from the Journal Editor (Pavel Trofimovich): This article is published with special permission from the board of directors of *Language Learning*, following regular peer review by four reviewers. The study emerged in conjunction with work on the multisite replication reported by Morgan-Short et al. (2018), funded by a *Language Learning* research grant to Marsden and Morgan-Short. That financial support was applied for and received before Marsden and Morgan-Short joined the editorial team of *Language Learning*.

This systematic review was presented in two colloquia on replication convened by the first two authors at the annual conferences of the American Association for Applied Linguistics (Portland, OR, March 2017) and the European Second Language Association (Reading, UK, August 2017). We thank our copresenters and the audiences at those colloquia for insightful discussion and feedback. We are also very grateful to Luke Plonsky for helpful advice during the initial stages of this systematic review. Some partial financial support was provided by the British Academy award ARP AQ160001. An earlier version of a small subsection of this synthesis, partially funded by the UK Economic and Social Research Council (RES-062-23-2946), was presented at The Instruments for Research into Second Languages (IRIS) Conference, University of York (September 2013) and at The International Symposium on Bilingualism, Nanyang Technological University, Singapore (June 2013).

This article has been awarded Open Materials and Open Data badges. All materials and data are publicly accessible via the IRIS Repository at https://www.iris-database.org/iris/app/home/detail?id=york:934328. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

Correspondence concerning this article should be addressed to Emma Marsden, Centre for Research into Language Learning and Use, Department of Education, University of York, York, YO10 5DD, United Kingdom. E-mail: emma.marsden@york.ac.uk

DOI: 10.1111/lang.12286

Despite its critical role for the development of the field, little is known about replication in second language (L2) research. To better understand replication practice, we first provide a narrative review of challenges related to replication, drawing on recent developments in psychology. This discussion frames and motivates a systematic review, building on syntheses of replication in psychology, education, and L2 research. We coded 67 self-labeled L2 replication studies found across 26 journals for 136 characteristics. We estimated a mean rate of 1 published replication study for every 400 articles, with a mean of 6.64 years between initial and replication studies and a mean of 117 citations of the initial study before a replication was published. Replication studies had an annual mean of 7.3 citations, much higher than averages in linguistics and education. Overlap in authorship between initial and replication studies and the availability of the initial materials both increased the likelihood of a replication supporting the initial findings. Our sample contained no direct (exact) replication attempts, and changes made to initial studies were numerous and wide ranging, which likely obscured, if not undermined, the interpretability of replication studies. To improve the amount and quality of L2 replication research, we propose 16 recommendations relating to rationale, nomenclature, design, infrastructure, and incentivization for collaboration and publication.

Keywords replication; methodology; systematic review; research design; publishing; second language

Introduction

Replication studies are considered by many to play a fundamental role in any scientific endeavor. When using the same materials and procedures as a previous study, replication studies serve to test the reliability of the previous study's findings. When altering specific methodological or participant characteristics of a previous study, they serve to test generalizability of the earlier findings under different conditions. One indication of the importance of replication is found in the 50 or more calls for replication research in the field of second language (L2) research alone (see references for 50 calls and commentaries in Appendix S1 in the Supporting Information online): from Santos (1989) through Polio and Gass (1997) to very recent proposals for specific replication studies, such as Vandergrift and Cross (2017) and even a book-length treatment (Porte, 2012). Beyond these calls, efforts to actively promote and facilitate replication studies have also emerged. For example, the Instruments for Research into Second Languages (IRIS) repository (http://www.iris-database.org) was established in 2011 and holds, at the time of writing, over 3,800 materials that can be used for replication, among other purposes, in L2 research (Marsden & Mackey, 2014; Marsden, Mackey, & Plonsky, 2016). The Open Science Framework (https://osf.io), also established in 2011, provides a web

infrastructure to facilitate collaboration and has been used for large replication efforts in psychology (e.g., Open Science Collaboration, 2015), which continue to make waves in academia (Laws, 2016; Lindsay, 2015; Martin & Clarke, 2017) and the general media (Baker, 2015; Devlin, 2016). In some fields, a flourishing metascience, that is, the scientific study of science (see Munafò et al., 2017), has included syntheses assessing the quantity and nature of replication efforts, for example, in education (Makel & Plucker, 2014) and in psychology (Makel et al., 2012).

The driving force behind this battery of calls, commentaries, infrastructure, and metascience is a perceived crisis in the state of replication research. The severe concerns underpinning the alleged crisis have several dimensions relating to: (a) the (small) amount of published replication research; (b) the (poor) quality of replication research; and (c) the (lack of) reproducibility, which refers to the extent to which findings can(not) be reproduced in replication attempts that have been undertaken. These concerns speak to the very core of science, raising fundamental questions about the validity and reliability of our work. Indeed, some commentators have called replication the "gold standard" of research evidence (Jasny, Chin, Chong, & Vignieri, 2011, p. 1225) and a "linchpin of the scientific process" (Let's replicate, 2006, p. 330).

In the field of L2 research, given the importance of replication and the 50 calls for replication in L2 research that we identified, one might expect a substantial number of published replication studies by now. However, a perceived lack of prestige, excitement, and originality of replication plagues L2 research (Porte, 2012), as it does other disciplines (Berez-Kroeker et al., 2017; Branco, Cohen, Vossen, Ide, & Calzolari, 2017; Chambers, 2017; Schmidt, 2009), and these perceptions are thought to have caused, at least in part (directly or indirectly), alleged low rates and a poor quality of published replication studies. However, a systematic metascience on replication research has not yet been established in the field of L2 research, leaving a poor understanding of the actual number and nature of replication studies that have been published.

The current study begins to address this gap through narrative and systematic reviews. The narrative review considers challenges in replication research and is largely informed by commentaries and metascience from psychology, given that the cognitive and social subdomains of psychology are highly influential in L2 research, and also from education, another key sister discipline. The narrative review is organized around four broad themes: (a) the quantity of replication research, (b) the nature of replication research, (c) the relationship between initial and replication studies, and (d) the interpretation and extent of reproducibility of the findings of initial studies. To gain insight

into these issues in the context of L2 research, the systematic review provides a synthesis of L2 studies in journal articles that self-labeled as replications. The research questions and methods of the systematic review were largely determined by the narrative review but also emerged through the design and piloting of the coding instrument. Finally, we offer further discussion and 16 recommendations for future replication work that draw on our narrative and systematic reviews and on our experience of carrying out multisite (Morgan-Short et al., 2018)[1] and single site (Faretta-Stutenberg & Morgan-Short, 2011; Marsden, Williams, & Liu, 2013; McManus & Marsden, 2017; Morgan-Short, Heil, Botero-Moriaty, & Ebert, 2012) replications. We start from the widely agreed premise that testing the reproducibility of findings should have an essential role in the testing and refinement of theory, at least for hypothesis-testing epistemologies that seek to ascertain generalizability and for other epistemologies in which constructs are deemed to be definable and observable. Thus, our overall aim is to provide conceptual clarification and an empirical base for future discussion and production of replication studies, with a view to improving the amount and quality of L2 replication research.

Narrative Review of Concerns and Challenges Related to Replication

The primary aim of this narrative review is to consider key issues related to replication research and to indicate how aspects of the narrative review inform the aims, scope, structure, and methods of our systematic review. First, we clarify our use of the terms *replicable*/*replicability* and *reproducible*/*reproducibility*, given some debate surrounding these terms (National Academies of Sciences, Engineering, and Medicine, 2016; National Science Foundation, 2015). The term replicable/replicability commonly serves two functions, and we have tried to ensure at each use whether we refer to either (a) the extent to which it is possible to carry out a study again (e.g., whether sufficient information and materials are available to allow replication of the study itself, also known as repeatability) or (b) the extent to which the results of a replication study are similar to those of the initial study (i.e., replication of findings). The term reproducible/reproducibility is used in a more marked way to refer only to (b), in line with the recent developments in the field of psychology (e.g., Open Science Collaboration, 2015).

The Quantity of Replication Research

To understand the state of replication research in a particular field, one must first determine the quantity of replication research that has been undertaken.

To do this, those studies which should be counted as replication research must be identified. This is not a trivial matter. Given a broad definition (e.g., studies investigating related questions using similar designs and materials), a very large number of studies could be called replications (see Plonsky, 2012, for discussion of the extent to which studies included in a meta-analysis could be considered replications, and VanPatten, 2002a and 2002b, for narrower conceptualizations). On the other hand, even studies that fall into a narrower definition of replication (e.g., investigating the same research questions with a design and materials that are as similar as possible to an earlier study) may not label themselves as replications. To illustrate, of the four studies that were part of the replication sequence extended by Morgan-Short et al. (2018), only one (Wong, 2001) turned up in our systematic review as a self-labeled replication. Given this subjectivity and inconsistency and, more importantly, given that we wanted in our systematic review to ascertain the extent to which the term replication has been used to label studies reported in journals, we used instead the self-identification of authors, that is, studies that self-labeled in the title or abstract as a replication study. This is similar to the approach of Makel et al. (2012) and Makel and Plucker (2014), who examined the state of replication in psychology and education, respectively, and avoided the need to create customized definitions of replication studies. However, we acknowledge that this approach does not encompass all research that could be viewed as a replication, an issue considered more fully in our recommendations for the field. Throughout this article, we use the term replication to refer to a replication study, that is, one that attempted, to some degree, to replicate a previous study's aims and methods. Our use of the term replication alone makes no allusion to whether the study succeeded (or indeed aimed) to replicate the methods exactly nor to the extent to which earlier findings were reproduced.

In addition to identifying replications, we must consider the nomenclature of subtypes of replication. In the field of psychology, an early proposal of three subtypes was made by Lykken (1968): (a) literal replication, in which additional participants were recruited to the same study; (b) operational replication, which used the same methods and conditions; and (c) constructive replication, where the claimed relation between constructs was tested using any methods the replicator wished. Others have converged on two subtypes (Makel et al., 2012; Schmidt, 2009): (a) direct replication, where there are no intentional or significant alterations of the initial study, considered "the means of establishing reproducibility of a finding with new data" (Open Science Collaboration, 2015, p. 1), and (b) conceptual replication, where there is intentional adaptation of the initial study to investigate generalizability to new conditions,

contexts, or study characteristics. Using this distinction, Makel et al. found that 81.9% of replications in psychology were conceptual, 4.1% were categorized as both conceptual and direct, and 14% were direct. The latter figure is most likely considerably higher now given the recent surge of direct replications (see below).

One problem with such dichotomous labeling is that for conceptual replications the number and type of changes to the initial study can vary and/or be vague, making it difficult to assess whether a study can test the effects of new constructs or of boundary conditions (i.e., study features that help determine the limits of generalizability to, for example, participants with different characteristics from those tested in the initial study). Indeed, Earp and Trafimow (2015) have provided a framework for conceptualizing different types of replications falling along a multidimensional spectrum, with each type serving a different purpose.

In L2 research, issues of nomenclature for different types of replications have also been a source of confusion (Polio, 2012b). Porte (2012) provided a taxonomy of three broad types of replication: (a) exact or literal; (b) partial, approximate, or systematic; and (c) conceptual or constructive. However, the extent to which this recommendation has been adopted by the field in a systematic manner remains unclear. Thus, our synthesis aimed to examine the nomenclature used for self-labeled replication research and the extent to which different labels have reflected the number and types of change between initial and replication studies. With this insight, we go on to propose a clear and principled nomenclature for the field.

On a final note about nomenclature, in the current reviews, we have used the term *initial study* rather than *original study* when referring to studies that were replicated. This is because studies are rarely if ever truly original in the sense of being a completely novel idea. Also, original carries negative connotations for its replication because it could imply that anything that is not original cannot share other characteristics broadly associated with originality, such as being innovative, fundamental, or agenda setting.

After replications have been identified and classified by type, issues of quantity can then be examined. In the field of education, Makel and Plucker (2014) found a publication rate of 0.13% for replication studies (221 out of 164,589 articles) in the 100 highest-impact journals between 1938 and 2014. In the field of psychology, Makel et al. (2012) estimated that among the top 100 journals between 1900 and 2010, the replication study publication rate was 1.07%, though this rate is now likely to be higher given recent multiple, direct replication projects: the Many Labs project (Klein et al., 2014), the Pipeline

Project (Schweinsberg et al., 2016), the Registered Reports project (Nosek & Lakens, 2014), and the Reproducibility Project (Open Science Collaboration, 2015). In business, marketing, and communication journals, replication rates have ranged from 1 to 3% (Evanschitzky, Baumgarth, Hubbard, & Armstrong, 2007; Hubbard & Armstrong, 1994; Kelly, Chase, & Tucker, 1979). In the field of L2 research, the rate of replication studies is perceived as being low, but without systematic data on this, concerns to date have necessarily been speculative.

Attempts to improve rates of replication have met many challenges (Porte, 2012), including some imposed by publishing venues themselves. The quantity of replication is, perhaps, influenced by the extent to which journals encourage or discourage replication. To investigate how psychology journals approach this issue, Martin and Clarke (2017) reviewed the scope sections of author guidelines of 1,151 journals and found that 63% did not state that they accepted replications, but neither did they discourage them; 33% implicitly discouraged them by emphasizing originality, novelty, or innovation of submissions; 3% of journals stated that they accepted them; and 1% actively discouraged replications by stating that they did not publish them. The fact that only 3% of journals stated that they accepted replications may partly be due to the perceived impact, and hence prestige, of replication. However, this perception may not reflect reality. To illustrate this with an example from the field of education research, Makel and Plucker (2014) found that the median citation count of replications was 5 (*range* = 0–135), compared to 31 for the initial studies (*range* = 1–7,644). However, this difference is not surprising, as initial studies have more time to be cited and high citation counts are often the reason for replicating them in the first place. Furthermore, as Makel and Plucker note, five citations for replications is relatively high, given that only one of the top 100 education journals had a 5-year impact factor higher than 5. For the field of psychology, Makel et al. (2012) found that the median number of citations of replications was 17 (*range* = 0–409), compared to the mean of 64.5 of initial studies (*range* = 1–2,099), and this was also observed as being relatively high given that only three of the 100 analyzed journals had a 5-year impact factor greater than 17. Thus, contrary to expectations, replications may have had a higher impact than the average article in their field as represented by journal impact factors.

Motivated and informed by previous work that quantifies replication in psychology and education, our systematic review had two key purposes: (a) to shed light on the quantity of replication in L2 research, for which we calculated the rate of replication, examined which journals have published replications, and documented whether journals discourage or encourage them, and (b) to

estimate the impact of published replications, for which we investigated the number of citations of replications and the impact factor of the journals that publish them.

The Characteristics of Research Studies that Warrant and Lead to Replication

Beyond questions about the quantity of replication, we considered the kind of research that the field appears to support as meriting replication. The extent to which reproducible findings are deemed to be a desirable ambition can vary according to different ontological, epistemological, and methodological perspectives (Markee, 2017; Polio, 2012a; Porte, 2012; Porte & Richards, 2012). There is a high degree of consensus that replication, particularly when narrowly defined as direct or close replication, is not appropriate or useful for all types and stages of research (e.g., ideological or interpretative approaches, exploratory or grounded research, or case studies). There is also clear consensus that replication is of value for a large portion of research, usually that which involves some hypothesis testing and/or data that are quantitative (either at the collection or coding stage). This may be because materials, measurements, and analyses are designed to be reproducible for this type of research so as to ensure generalizability, which conforms to the epistemologies of such research. Putting those relatively well-rehearsed issues aside and focusing mainly on the large body of research in which the desirability of replication is rarely controversial, a variety of suggestions have been made about characteristics of research that warrant replication endeavors, for example, the significance and design of the initial study.

Regarding the significance of an initial study, Nosek and Lakens (2013) suggested that "important" research to replicate is that which is "often cited, a topic of intense scholarly or public interest, a challenge to established theories, but [it] should also have uncertain truth value (e.g., few confirmations, imprecise estimates of effect sizes)" (p. 59). Thus, the number of citations may be a warrant for replication. For example, Makel et al. (2012) suggested that it would be surprising if replications had not been triggered after (an admittedly arbitrary) 100 citations of a study.

However, citation counts alone are unlikely to offer reliable or sufficient motivation for replication. Importance also stems from the research community's views on what research needs to be replicated to inform theory, method, or practice. We briefly mention four possible approaches for establishing what is important. The journal *Language Teaching* (*LT*) includes an article type in which authors justify and describe specific replications that should be done,

and indeed, 12 such articles had been published at the time of writing (see Appendix S1 in the Supporting Information online). However, the extent to which this unique initiative leads to replication is unknown. Somewhat surprisingly, in our study sample (described below), we found no published replications that followed the suggestions made, nor did we observe a general increase in the number of replications published after these article types were introduced in 2014 (with Basturkmen, 2014).

Another approach is to crowdsource proposals for replication (see PsychFileDrawer http://www.psychfiledrawer.org/top-20), whereby a social media platform allows people to propose and vote on the studies that they would like to see replicated. Since it began in 2012, this archive of replications currently holds 71 reports, but the extent to which such an initiative, which is outside the standard publication venues, will have a lasting impact on the number or quality of replications is unclear. Another possibility is for journal editors to invite replications of particular studies, as is occasionally done by the editors of the Registered Replication Reports in *Perspectives in Psychological Science*. This approach exerts strong editorial influence over the types of studies that are replicated and how they are replicated and demands a heavy editorial role (D. Simons, personal communication, September 16, 2016). A final possibility is that researchers themselves provide theoretical and methodological justifications in the rationales sections of their replication studies, and these arguments are evaluated via current peer-review mechanisms. All these approaches may help to establish which research merits replication, but data are needed to ascertain the extent to which they are effective mechanisms for improving the amount, quality, or perceived prestige of replications.

Another factor potentially indicating importance, and thus a need for replication, are "surprising" findings (see Makel et al., 2012, p. 540; Porte, 2012, p. 7). Surprising could be, for example, large effect sizes when a meta-analysis would predict them to be smaller (or vice versa). Laws (2016) described all of the 13 studies replicated in Nosek and Lakens's (2014) Special Issue as "curios" (p. 2), with odd findings. Interestingly, 10 of those 13 initial findings were not reproduced. Thus, one (arguably undesirable) downside to surprising findings serving as a rationale for replication is that if the rate of reproducing findings from such research is unusually low, the overall rate of reproducibility for a field may appear to be lower than it actually is (Hartshorne & Schachner, 2012; Laws, 2016). Also, using the surprising-findings rationale alone as a warrant for replication could introduce a type of reverse publication bias, whereby finding no effect in a replication (where an effect or statistical significance was found

in the initial study) is considered the more publishable and citable outcome (Ioannidis, 2005; Luijendijk & Koolman, 2012).

Finally, the statistical significance of a study's results may have (undue) influence on its perceived importance for replication. Publication bias—a tendency for journals to publish and/or researchers to submit only statistically significant findings—is a widely acknowledged problem, and null findings are confined to the "file drawer," a term coined by Rosenthal (1979) and a phenomenon documented by many scholars (e.g., Bakker, van Dijk, & Wicherts, 2012; Schmidt & Oh, 2016; Sterling, Rosenbaum, & Weinkam, 1995; Sutton, 2009). Though the extent of field-wide publication bias in L2 research has not yet been systematically studied, it likely exists (Fanelli, 2012; Plonsky, 2013), and several meta-analysts have found some evidence of it in specific domains (Lee & Huang, 2008; Lee, Jang, & Plonsky, 2015; Plonsky, 2011). This means that even unintentionally, anyone choosing a study to replicate is likely, due to chance alone, to select one with statistically significant findings. To give one example of this phenomenon, Laws (2016) noted that the four multisite replications that he reviewed almost entirely neglected null findings. Similarly, in the Open Science Collaboration (2015) project, only three of its 100 initial findings were null. Yet it is of course useful to carry out replications of studies with null or borderline findings. For instance, for the three null studies replicated by the Open Science Collaboration, the replications confirmed two as null but produced statistically significant findings for the other one (see also Morgan-Short et al., 2018).

The need to replicate studies with null findings is particularly important in L2 research, where sample sizes are often too underpowered to reject the null hypothesis with an average post hoc power of .57 (Plonsky, 2013), the statistical equivalent of "tossing a coin in the air and hoping for heads" (Plonsky, 2015, p. 29). In sum, the absence of statistical significance in an initial study may: (a) not validly indicate the absence of an effect but rather be an artefact of other issues, such as small sample size or chance findings; (b) be a theoretically or practically useful finding that does merit corroboration via replication; and (c) lead to dichotomous rather than nuanced interpretations. Thus, statistical significance alone serves as a dubious warrant for replication.

Beyond the significance of an initial study, a warrant for replication must also consider research design. Indeed, suggestions have been made for researchers to select studies to replicate based on a set of problematic characteristics and findings. For example, Lindsay (2015) proposed that researchers be on the "lookout for this troubling trio: (a) low statistical power, (b) a surprising result, and (c) a p value only slightly less than .05" (pp. 1827–1828).

Another proposal—a quantitative doping test for science proposed by Schimmack (2016)—is known as the replicability index (R-index) and is used to evaluate the statistical replicability of a set of studies. It is based on the difference between median estimated power and likely rate of reproducing findings, which results in the so-called inflation rate. Results of studies with these characteristics that can cause concern may be due to questionable research practices, such as not reporting all outcome measures or conditions, only reporting statistical tests that found statistical significance, data peeking before deciding when to stop testing participants or whether to exclude (particular definitions of) outliers, and HARKing—hypothesizing after the results are known (Chambers, 2017; Kerr, 1998; Lindsay, 2015). Thus, replication could help ascertain the likelihood of findings being actually valid or merely an artefact of such issues.

Even if a replication is warranted, other design characteristics of studies may affect the feasibility of carrying out a replication. Practicalities of time and resources may impede the replication of certain studies, meaning that studies termed cheap and easy by Laws (2016) are replicated while replication in some subdomains is "likely to remain castles in the air" (p. 3). One likely manifestation of these practical constraints was the Many Labs Replication Project (Klein et al., 2014), which delivered a single 15-minute questionnaire (combining 13 earlier experiments) to 6,344 participants across 12 countries via 36 research groups. In L2 research, designs that are usually more costly involve longitudinal designs (e.g., experiments with pre-, post-, and delayed posttests as opposed to one-shot or cross-sectional designs), one-to-one measures (e.g., oral production tests versus group-delivered pen-and-paper or computer-based tests), equipment that is expensive to purchase or utilize (e.g., eye-tracking or neuroimaging hardware), and participant populations that are difficult to reach (e.g., rarer language combinations, schools, heritage speakers, or participants linked to a specific history or culture). Replications with such designs may be underrepresented compared to more easily administered designs.

Another key characteristic that affects whether, and how well, a replication study can be carried out is the transparency of the initial research because availability of materials and data, as well as thorough reporting, are needed for replication and are particularly important for independent or direct and partial replications. For example, the availability of data helps replicability and the evaluation of reproducibility because researchers can (a) increase the sample size of previous research; (b) combine their data with previous data in new analyses; (c) reanalyze data to assess the reliability of the initial analyses (which is specifically termed reproducibility by National Science Foundation, 2015); and (d) evaluate the parity of samples, which is particularly critical in

L2 research as participant demographics, such as proficiency, age, and first language (L1), are known to affect language development.

However, an academic culture in which there is little chance of replication happening or being published reduces the perceived need to make research replicable through materials and data availability and transparent reporting because researchers might very reasonably ask themselves, "Is anyone really going to attempt to replicate this?" This no doubt partially accounts for a history of inadequate reporting practices (e.g., as noted by Derrick, 2016; Han, 2016; Larson-Hall & Plonsky, 2015; Plonsky & Derrick, 2016), poor transparency of materials (Marsden & Mackey, 2014; Marsden et al., 2016; Marsden, Thompson, & Plonsky, in press), and very scarce availability of data (Larson-Hall & Plonsky, 2015; Larson-Hall, 2017; Plonsky, Egbert, & LaFlair, 2015). For an overview of these issues, see Marsden (in press); for discussions of similar challenges in linguistics, see Berez-Kroeker et al. (2017), and in psychology, see Fecher, Friesike, and Hebing (2015), Lindsay (2017), and Wicherts, Borsboom, Kats, and Molenaar (2006). Indeed, aiming to address this situation, the Transparency and Openness Promotion (TOP) Guidelines (Nosek et al., 2015a, 2015b) encourage journals to incentivize/require their authors to make their materials and data transparent. These guidelines also set explicit benchmarks about the levels to which journals promote replication (discussed further below), thus drawing clear links between replication and the transparency of materials and data.

In sum, issues such as the initial reporting of methods, results, and analysis; the availability of the initial materials and data; and the resources needed may all reduce the likelihood, quality, or usefulness of replication (even when a replication is clearly warranted). Motivated by these issues, in our synthesis we probed the question of what warrants and leads to replication by examining the following characteristics of studies that have been replicated in L2 research: (a) citation counts; (b) broad findings (statistically significant or null); (c) designs, measures, and sample sizes (to investigate the extent to which replication has been concentrated on cheap and easy designs); (d) transparency of reporting; and (e) availability of materials and data.

Extent of Change Between Initial and Replication Studies

The rationale for replicating a study can also be determined by the nature of the specific changes made to the designs of the initial studies. Many researchers include caveats about their studies, suggesting that future research should replicate the study to test boundary conditions, that is, the extent of generalizability to, for example, a different outcome measure, experimental

design, L1 background, modality, target language, or age or proficiency of participants. However, making many or unacknowledged/unspecified changes to a study lies in tension with being able to account for whether differences in findings compared to the initial study are ascribable to the heterogeneity that was introduced (intentionally or otherwise) or to some other factor. This issue was tackled by Klein et al. (2014) in their direct replications, wherein heterogeneity between initial studies and replications was kept to a minimum except for two key variables (participant nationality, lab vs. online delivery). These researchers estimated the proportion of variation in effect sizes attributable to heterogeneity of implementation rather than to chance, showing that the effects of heterogeneity in those variables were nonexistent or very small in most cases.

A related issue is that even when maximum effort is made to maintain homogeneity of implementation between initial and replication studies, there may be auxiliary assumptions embedded in the hypotheses or design of the initial studies. Regardless of whether these assumptions are well understood or not, if the replication study violates them inadvertently, this can affect the outcomes and could result in findings that do not align with those of the initial study, as discussed by Trafimow and Earp (2016). As a preliminary investigation into the extent and nature of heterogeneity in L2 replication research, in the current synthesis, we sought to collect data on the types of changes that have been made in replications and the extent to which heterogeneity between initial and replication studies was intentional (for partial or conceptual replications), explicitly acknowledged (for all types of replication), or not acknowledged by the authors.

Another common caveat in the concluding sections of articles is that replication is required due to the small sample size of the study. It might therefore be expected that self-labeled replications have a larger sample size than initial studies. However, a survey by Tversky and Kahneman (1971) found that most social scientists believed that if a finding had been observed with a certain sample size, the same outcome should be observed with a smaller sample. Given a scenario in which an initial study (e.g., $N = 40$) produced statistically significant findings and a replication (e.g., $N = 30$) did not, most respondents gave an explanation for this difference related to theory, measurement constructs, or participant characteristics rather than an explanation related to, more simply, the higher power of the initial study. To eliminate low power as a potential explanation of nonreproduced results, the sample size of a replication study should be at least the same as the sample of an initial study. Furthermore, it may be desirable for a replication to have a larger sample size. Earp, Everett, Madva, and Hamlin (2014) argued that publication bias and the concomitant

issue of increased likelihood of results being statistically significant and/or effects being at the high end of the distribution can mean that the same sample size might fail to reproduce the earlier findings or detect an effect at all. Even with a larger sample size, a replication study may not have sufficient power to find an effect similar to that of the initial study (or any meaningful effect) if that effect was spurious or overinflated. A priori power analyses, at a minimum, can help to address the issue of whether a change in sample size is needed for a replication (for related discussion, see Simonsohn, 2016).

To investigate the heterogeneity and sample size issues discussed in regard to replication, we documented the nature and number of changes between the initial and replication studies. We also explored whether these changes were associated with the nomenclature of replications (e.g., direct vs. conceptual) and with the extent to which their findings supported the initial studies.

Extent of Reproducibility

The extent to which replications demonstrate reproducibility of earlier findings partly depends on how the term "reproduced" is defined. When reproducibility has been quantified in syntheses and meta-analyses of replication in other fields, there has been a range of outcomes. For direct replications in psychology, the Many Labs project found that 10 out of 13 replications reproduced the initial findings, whereas the Registered Reports project (Nosek & Lakens, 2014) found that 10 out of 13 did not; meanwhile, four high-powered replications by Rohrer, Pashler, and Harris (2015) found no support for earlier studies. The Open Science Collaboration (2015) used different measures of reproducibility for their direct replications and found that, based on null hypothesis significance testing (NHST), only 36% of replications yielded significant results compared to 97% of the initial studies. However, NHST can only provide a dichotomous perspective—significant or nonsignificant (e.g., Norris & Ortega, 2000; Norris, Plonsky, Ross, & Schoonen, 2015)—and does not allow for a more fine-grained measurement of the extent of reproducibility. Broader categories for assessing reproducibility are needed to provide a more tolerant, less rigid measure that reflects some of the variability inherent in many studies, particularly likely in research with human participants and/or multiple complex variables (for discussion, see Earp, 2016, and Trafimow & Earp, 2017).

Another approach is to use subjective ratings of reproducibility. Interestingly, however, the subjective ratings approach of the Open Science Collaboration (2015) led to assessments of reproducibility that were very similar to their NHST approach. Based on 7-point subjective ratings ranging from *virtually identical findings* to *not at all similar*, 39% of replications were deemed to

have reproduced the initial result, compared to 36% according to NHST. Perhaps the similarity in the findings emerged because subjective ratings may have largely relied on the NHST reported in the studies. However, different outcomes were found when using effect sizes to assess reproducibility: Reproducibility increased to 47% when it was based on whether an effect size fell within the 95% confidence interval (CI) of the initial effect size. Finally, using yet another measure of reproducibility, the reanalysis of the Open Science Collaboration (2015) data by Patil, Peng, and Leek (2016) found that 77% of the effect sizes were within a 95% prediction interval of the initial effect size (see Francis, 2012; Lindsay, 2015; Maxwell, Lau, & Howard, 2015; and Stroebe & Strack, 2014, for further discussion of ascertaining reproducibility; and Marsman et al., 2017, for Bayesian approaches to assessing reproducibility).

Other, broader syntheses of the replication effort within whole disciplines have made estimates of the extent to which findings have been reproduced, as reported by the authors themselves, using subjective rating measures. As in our systematic review, this is a suitable estimate mechanism given that the replications included in these syntheses were not direct, and so a precise, quantitative assessment of reproducibility was not a key aim. In the field of education, Makel and Plucker (2014) used a subjective three-level scale to rate reported replication success in existing replications, of which only 14% were direct. They found that 67.4% of replications reported successfully replicating the initial findings, 19.5% replicated some but not all findings, and 13.1% failed to replicate the initial findings. Using a similar rating scale, Makel et al. (2012) found that 78.9% of studies successfully reproduced the initial findings, 9.6% did not, and 11.4% reported mixed support. Overall, the reproducibility rate in these fields has been calculated to range from around 36% to 79%, but that rate has depended on how it was assessed, among several other factors.

One such factor is that reproducibility is likely to vary according to subdomain. For example, in the Reproducibility Project, 25% of effects in social psychology were replicated (according to the $p < .05$ criterion), compared to 50% of effects in cognitive psychology; however, as noted above, there are problems with using the dichotomous and arbitrary cutoffs of NHST. A second factor may be the type of replication. For direct replications, where minor differences in implementation are not theorized to influence the findings, expectations for reproducibility are high. Although it cannot be expected that all direct replications would find the same magnitude of effects or patterns of statistical significance as their initial studies (Francis, 2012; Laws, 2016; Lindsay, 2015; Open Science Collaboration, 2015), one might predict effect sizes within

the 95% CIs of the initial effect sizes, and (at the very least) the same direction of differences or associations. On the other hand, for partial and conceptual replications, which intentionally introduce change to initial study designs, researchers may make theoretical predictions about why the change may (or may not) make a difference to findings. That is, partial and conceptual replications introduce more than just incidental operational heterogeneity, sometimes with the expectation of not reproducing the initial findings. An example of this from our sample of replication research is Ellis and Sagarra (2011), who intentionally introduced more verb inflectional diversity into their materials and found the difference compared to the findings of the initial study that they were expecting to find. However, the intuitive expectation of less supportive findings emerging from partial or conceptual replication studies, compared to direct replications, does not seem to be observed consistently. For example, Makel et al. (2012) found that, in fact, conceptual replications supported initial findings at a descriptively higher rate than direct replications (82.8% vs. 72.9%), whereas Makel and Plucker (2014) found the reverse (66% vs. 71.4%). However, neither pattern was statistically significant. In light of these issues, in the current synthesis, we avoided describing replications as failed or unsuccessful. Given that our sample did not yield any direct replications, not reproducing findings (however that is determined) does not necessarily indicate flaws in either the initial or replication studies, as it could in fact have been expected. That is, we did not set out to evaluate the overall level of reproducibility in the field as being good or bad.

A third factor in reproducibility may lie in the independence of the replication researchers in relation to the initial researchers. In education, Makel and Plucker (2014) found that nearly half (48.2%) of the replications were conducted by the same research team who had published the initial research. When at least one author was on both the initial and replication articles, 88.7% of replications supported the initial findings, although the rate dropped to 70.6% if the replication was published in a different journal. With no author overlap, the rate dropped further, with 54% of replications supporting initial findings. In psychology, Makel et al. (2012) found 91.7% supported initial findings when there was author overlap, versus 64.6% when there was no overlap. Given the high rate of reproducibility with author overlap, Koole and Lakens (2012) focused only on independent replications in their set of recommendations for replication, arguing that "the most compelling direct replications are conducted independently by different researchers than the original study" (p. 609). This was a key motivator for the preregistered multisite replications published by *Perspectives in Psychological Science* (soon to move to *Advances in Methods*

and Practices in Psychological Science), in which research teams all have access to the same materials but conduct the study independently (and in some cases, do not look at the data until they have been passed to the replication convener or coordinating editor).

Independence of researchers does not necessarily reduce bias, as bias can also affect an independent replicator, who may predict findings against others' work (Bakan, 1967). In fact, author overlap may bring perceived advantages. In a climate where there is little sharing of materials and data, author overlap may increase the chances of better fidelity to the initial study's materials and protocols. Indeed, Makel et al. (2012) found that most direct replications were conducted by authors of the original studies. Similar to the availability of data being associated with better reporting and stronger evidence (Wicherts, Bakker, & Molenaar, 2011), the availability of instruments may affect the nature of results too, for example, by increasing the likelihood of demonstrating support for the initial study's findings. In our own synthesis, we explored this possibility, partly driven by a concern that although more supportive findings may be a perceived benefit of author overlap, this may not be beneficial for the speed and objectivity of the broader scientific endeavor, as giving others to access materials may facilitate faster and, perhaps, less partisan replication efforts.

The current synthesis did not aim to evaluate the reproducibility of L2 research. This decision was determined partly by the fact that we found no direct replications and observed widespread intentional and unintentional heterogeneity between initial and replication studies and partly by the need to limit the size of our undertaking.[2] However, we do provide a preliminary examination of whether the extent to which replications supported the initial findings, as claimed by the replicating authors, was associated with certain factors, such as the subtype of replication, the independence of researchers, and the availability of materials. This examination is based on subjective ratings targeting the extent to which the replications' findings were reported as supporting the initial findings, as used by Makel et al. (2012) and Makel and Plucker (2014). Therefore, this analysis relied on how the replicating researchers presented and discussed their data and analysis in relation to the earlier study.

A Systematic Review of Self-Labeled Replication

Aims

The above narrative review of commentaries, meta-analyses, and metascience on replication closely informed the research questions and methods for our systematic review of replication in L2 research. For example, the syntheses of replication by Makel et al. (2012) and Makel and Plucker (2014) in the fields

of psychology and education closely informed our investigations into (a) the quantity and nomenclature of replications, their publishing outlets, and citation counts; (b) relations between the authorship of replications and their initial studies; (c) the extent of independent replication (with/without authorship overlap, in same/different journals); and (d) whether findings were interpreted by authors as supporting or not supporting the initial studies. In these respects, our systematic review is, in broad terms, a conceptual replication of the systematic reviews conducted by Makel et al. (2012) and Makel and Plucker (2014), sharing common aims though with numerous differences in context and methods.

Other issues identified in our narrative review informed our systematic review, but had not, to our knowledge, been systematically examined in previous synthetic work on replication. For example, our narrative review of infrastructure and projects that have helped methodological transparency and collaboration in psychology led us to document the transparency of our L2 initial studies, such as their reporting and the availability of their materials, data, and analyses. This allowed us to examine the impact that methodological transparency and authorship overlap may have had on replication research, such as (a) whether and how replicators accessed materials and data, (b) the existence of interconnected series of initial and replication studies, and (c) associations between materials transparency and the extent to which replication findings supported the initial findings. Also, we wanted to estimate the time between a study and its replication when replications were published in articles separately from their initial studies (rather than within the same multiexperiment article, of which we found very few, in any case). Addressing these issues gave us insight into the procedural and cultural change that might be necessary to enhance the amount and quality of replication research.

Other aspects of our systematic review were also indirectly informed by the narrative review above but were sharpened a great deal during the process of doing the systematic review itself. For example, when our search did not yield any self-labeled direct replications, then documenting heterogeneity—the amount and nature of changes that had been introduced into the replication studies compared to the initial studies—became a major undertaking in coding the articles. This led us to examine whether the amount of these changes was related to self-labeling nomenclature and to the extent to which a replication supported the initial study's findings. Additionally, a small number of issues were incorporated into our review during the development of the coding scheme to document the kinds of studies that have been replicated in L2 research. These issues related to characteristics specific to L2 research, such as study design, measures, and participant characteristics.

In these ways, our systematic review converged on the following research questions:

1. How much self-labeled L2 replication has been published and in which journals?
   a. Which replication labels have been used?
   b. Which journals have published replications?
   c. What are the citation counts of replications, of their initial studies and of the journals in which the replication and initial studies were published?
   d. To what extent have closely interconnected series of initial and replication studies been conducted?
2. What kinds of L2 studies have been replicated?
   a. Have the findings from initial studies tended to be statistically significant or null?
   b. What were the designs and contexts of the initial studies?
   c. What were the participant characteristics in the initial studies?
   d. To what extent were the materials of the initial study accessible?
3. To what extent and how did researchers change the initial L2 studies?
   a. What are the overall extent and types of the changes?
   b. To what extent did the amount of change between initial and replication studies relate to nomenclature of replications?
4. To what extent did L2 replications support the findings of the initial studies?
   a. How did authors compare their findings with the initial findings?
   b. Which factors might have been associated with the extent to which replications supported initial findings: author overlap, amount of change from the initial study, transparency of the initial study's materials?

Methods for the Systematic Review

Searching

We focused our search on academic, peer-reviewed journals because we wanted to examine the extent of self-labeled replication in this medium, which has been identified as the primary channel for disseminating L2 research (Smith & Lafford, 2009; VanPatten & Williams, 2002). We therefore excluded replications in books, dissertations, conference proceedings, and the like, following procedures used in previous syntheses in the field (e.g., Plonsky & Gass, 2011; Plonsky, 2013). Admittedly, this left our sample susceptible to the effects of potential publication bias among journals. However, such bias would be a concern particularly for quantitative meta-analyses of substantive findings (because

effect sizes are likely skewed upward due to publication bias) but arguably less of a concern here because we did not undertake such a meta-analysis. Nevertheless, the file drawer problem is likely to affect replications as much as, if not more than, other studies due to concerns about manuscript rejection when findings do not align with those of the initial researchers (who might be chosen to peer review the manuscript).

First, our review of commentaries about L2 replication yielded six empirical replications for potential inclusion. We then performed a keyword search for articles in the Linguistics and Language Behavior Abstracts and PsycINFO databases that contained in their title or abstract the word *replicat** and either *second language* or *foreign language*, with no date restrictions. After we had combined results and removed duplicates, this yielded 891 hits (as of October 9, 2016). A Google Scholar search using these same keywords yielded a prohibitively high number of results (> 18,000); because we felt that our previous 891 hits provided a sufficiently representative picture for our purposes, these Google Scholar results were not used.

We then selected only articles that were in Social Science Citation Index journals and written in English. To be included in our review, articles had to present empirical research with data from L2 learners, educators, or materials. We had to exclude many of the articles found with *replicat** because they were false hits—researchers used the term to point to the need for replication of their own study or to claim their findings aligned with (replicated) earlier findings though the study itself was not a replication attempt (see also Makel et al., 2012, who found that only 68% of articles using *replic** were actual replications). Further details about exclusions, with examples, are available in Appendix S1 in the Supporting Information online. After implementing these exclusion criteria, we ultimately identified 67 replication articles and the 70 initial studies that they had replicated.

Coding the Studies

Our initial scheme, containing 61 categories for coding characteristics of these studies, was based on the narrative review above, including literature on replication in L2 research (e.g., Norris & Ortega, 2000; Polio & Gass, 1997; Porte, 2012). After 12 iterations during development, 42 of the original categories were maintained (marked ˆ in the coding sheet in Appendix S2 in the Supporting Information online); 19 of these 42 were modified slightly during coding development, and 94 categories were added such that the final coding scheme had 136 categories (marked # in the coding sheet in Appendix S2 in the Supporting Information online). Of these, 80 were categorical (27 dichotomous, 53

of which had three or more codes), 36 continuous, and 20 included open text. These categories captured information relating to seven clusters of characteristics consisting of:

- journal, article, and author information (25 categories),
- study design and participant characteristics (81 categories, including differences between initial and replication studies),
- analysis procedures (4 categories),
- findings (16 categories, including 14 relating to the nature of analysis and discussion of the two sets of findings),
- materials availability (4 categories),
- response/commentary from the initial author(s) (1 category), and
- additional notes (5 categories).

In the first pilot coding, all four of us coded two articles, then discussed our initial decisions, and changed the scheme accordingly. In the second pilot, two of us coded the same 14 replication–initial pairs of studies (21% of the total sample of studies). These were coded over several weeks, and the coding process included meetings with all four of us in which some aspects of the coding scheme were clarified and disagreements were addressed. Interrater reliability was calculated for the coding of these 14 pairs of studies. Of the coding categories, 57 allowed a Cohen's kappa reliability coefficient to be calculated whereas other coding categories (e.g., bibliographic information, long text answers, and entirely constant codes) could not yield a kappa value. The mean percent agreement between the two raters was 89%, and the mean kappa was .80. To set this in context, the reported kappa in other methodological syntheses has been .74 (Plonsky & Derrick, 2016), .56 (Plonsky, 2013), and .86 (Marsden et al., in press). To further enhance reliability of coding for the remaining studies, categories for which the percent agreement fell below 80% (13 columns) were reexamined by the two coders, who either amended or confirmed the initial codes. After this, the percent agreement for every category was at least 80%, and the mean interrater reliability was 94% (κ = .88). Using this finalized coding scheme, the two researchers individually coded the remaining 101 studies.

Analysis

Our analysis of the codes almost exclusively draws on descriptive statistics, such as percentages and measures of central tendency and dispersion because we sought to identify potential trends and formulate plausible accounts for them. During the analysis phase, nine columns were added to the coding sheet,

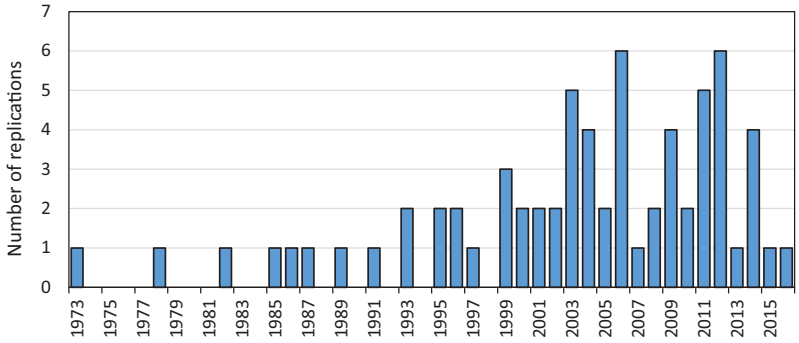

Figure 1 Self-labeled replications published in journals. [Color figure can be viewed at wileyonlinelibrary.com]

including those for article and journal citation data. The final coding sheet (including percent agreement rates, kappa values, and the data) is provided in Appendix S2 in the Supporting Information online and is openly available on IRIS (http://iris-database.org).

Results of the Systematic Review

Results are presented for each research question. Given the number and range of research questions, we provide some discussion with each set of results to render our responses more readable.

How Much Self-Labeled Replication Has Been Published and in Which Journals?

Our search found 67 self-labeled replications of 70 initial studies for a total of 129 study reports that were coded for further analysis.[3] All studies were published as journal articles except for five book chapters that were initial studies; three of these chapters were replicated in one replication study (Cobb, 2003). During our search, we also found 50 articles and chapters that were commentaries on or calls for replication in L2 research. This is just over two-thirds of the number of empirical self-labeled replications. In Appendix S1 in the Supporting Information online, replications are marked with *, initial studies with †, and commentaries and calls with °.

The earliest replication study was from 1973. A fairly steady increase began in the late 1990s until the most recent published replication appeared at the close of our search in October 2016 (Figure 1). In that time period, there was a mean of 1.55 ($SD = 1.69$) replications per year. The steady increase

Table 1 Terms used to label replications and articles self-labeling as replication in titles or abstracts ($k = 67$)

Terms used	Self-labeled
Close replication	1%
Approximate replication	3%
Partial replication	21%
Conceptual replication	4%
Replicat* (without a qualifier)	67%
Other	3%

probably reflects mainly an increase in volume of research rather than in the proportion of replications itself. However, some of the increase may be due to seminal papers promoting a synthetic research approach (i.e., an approach advocating consolidation and synthesis of research findings within and across various fields of inquiry), such as those of Polio and Gass (1997) and Norris and Ortega (2000), as well as dedicated replication article types in certain prominent L2 journals (e.g., launched in 1993 and refreshed in 2015 by *Studies in Second Language Acquisition* [*SSLA*] and started in 2014 by *LT*). The mean time between a study and its published replication was 6.64 years ($SD = 6.16$, $Mdn = 5$, $mode = 1$, $k = 11$, $range = 0$–37). This time delay demonstrates the need for a sustained infrastructure to help replications to be performed and published more quickly because ascertaining the generalizability and reliability of study findings can reduce the chance of self-perpetuating misinformed agendas and of drawing implications for practice too hastily (see also Koole & Lakens, 2012).

Which Replication Labels Have Been Used?

Examining the nomenclature used for replications, we found that after the single term replication, used in combination with extension in 25% of studies, the next most common label was partial replication in 21% of studies (see Table 1). However, a wide variety of terms were used, including strict replication, replication design, modified replication, and follow-up study. Many ($k = 24$) used multiple terms for the same study. Certain terms were never used despite having been used or recommended in commentaries on replication: true, direct, exact, quasi, and ceteris paribus. Overall, nomenclature was not precisely defined or consistent across studies, reflecting the confusion mentioned by Polio (2012b). We revisit nomenclature in our analysis of the extent to which labels reflected the amount of heterogeneity between initial and replication studies.

Table 2 Rates of replications in the five journals publishing the most replications (1973–2015)[a]

Journal characteristic	*SSLA*	*MLJ*	*LL*	*FLA*	*AP*	*M*	Total
Number of replications	11	8	5	5	4	6.6	33
Initial study in same journal	6	4	0	2	1	2.6	13
Total number of articles[b]	562	1,009	855	1,528	1,030	996.8	4,984
Replication rate	1.96%	0.79%	0.58%	0.33%	0.39%	0.81%	0.66%

Note. *SSLA* = *Studies in Second Language Acquisition*, *MLJ* = *The Modern Language Journal*, *LL* = *Language Learning*, *FLA* = *Foreign Language Annals*, *AP* = *Applied Psycholinguistics*. [a]2015 was the last complete year captured by our synthesis. [b]To calculate the denominator (total articles published), we used a start date of either 1973 (the date of our first replication) or the start of the journal if that fell after 1973.

Which Journals Have Published Replications?

Replication articles were found in 26 different journals. Five journals published four or more replications, giving a replication rate of 0.66% across these journals, calculated as the number of replication articles divided by the total number of research articles, excluding editorials and the like (see Table 2). Of the 26 journals that have published replications, the great majority of journals ($k = 21$) published three or fewer replications (see Table S3-1 in Appendix S3 in the Supporting Information online). Across all 26 journals, we estimated the replication rate as 0.26%. This was calculated as the number of replication articles divided by the estimated total number of articles. The estimated total number of articles was computed using the mean total of 996.8 articles produced by each of the top five journals in the time period found by the synthesis (see Table 2) multiplied by 26 journals, yielding a total of 25,917 articles. Expressed differently, the formula estimates that one in every 400 journal articles was a self-labeled replication. This is a generous estimate[4] of the rate of self-labeled L2 replication but it still falls below the mean rate in psychology in 2012, which would now be higher because of the recent surge in replications (as discussed above). We estimate that the field of L2 research may have a similar rate as education (calculated in 2014 at 0.13% by Makel & Plucker, 2014) or perhaps lower, given that the denominator for education used a much larger number of journals whereas we used only those journals that have published a self-labeled replication.

The low replication rate may be partially due to journals' (dis)encouragement of replications. Of the 26 journals that had published replications, only four explicitly stated that they accepted replications: *SSLA*,

Second Language Research (SLR), *LT*, and *Language Testing* (*LTest*). Interestingly, however, only one of these (*SSLA*) was in our top five list of journals publishing self-labeled replications, the others having published three, two, and two, respectively. Two of the four journals that stated they accepted replications emphasized originality in the first sentence of their aims/scope sections. Three of these four journals reserved specific strands for replications (*SSLA*, *LT*, *SLR*). Two of these strands were shorter article types, which might make overt comparability with initial studies difficult (we refer to this issue further in our recommendations about peer reviewing of replications). Ten of the 26 journals implicitly discouraged replications, with 9 of these emphasizing originality, novelty, or innovation in the first or second sentence of the aims/scope sections. Although three journals specified that methods should be clear enough to allow others to replicate the study, two of these did not explicitly state that they accepted replications in their own journal. Finally, two journals explicitly mentioned that null findings would not be grounds for rejection per se (*SLR* and *LTest*), both journals that encouraged originality and explicitly accepted replications).[5]

Beyond this analysis of the number of replications published by journals and the explicit and implicit messages that journals send to authors, it is not possible to determine how the replication rate of L2 research reflects the extent to which authors submit replications that are ultimately rejected versus the extent to which replications are simply not submitted. To obtain this information, surveys of editors and reviewers are necessary. The review by Martin and Clarke (2017) of such research showed that none has yet been done specific to language learning or education; data on this are central to improving our understanding of the causes of low rates of published replications.

What Are the Citation Counts Associated With Initial Studies and Their Replications?

With insight into the numbers and places of publication, we turn to examining the impact of self-labeled replications. Journal impact factors from the Web of Science (Thompson Reuters) and the total citations of the replication and initial studies (according to Google Scholar) were recorded in May 2017.[6] Table 3 shows that article citations were higher for initial compared to replication studies, which is unsurprising given that high citation often motivates replication and that initial studies had been available for citing over a longer period of time.[7] To take years since publication into account, we divided total citations by the number of years elapsed between publication and 2017 to provide mean citations per year. In terms of the relationship between median citations of replications and their initial studies, we found a ratio of 0.25 for L2 research,

Table 3 Article citation counts and journal impact factors for replication and initial studies

	Total article cites		Annual article cites		Journal 5-year IF	
Study type	*M* (*SD*)	*Mdn* (*range*)	*M* (*SD*)	*Mdn* (*range*)	*M* (*SD*)	*Mdn* (*range*)
Initial study[a]	364.03 (678.14)	173 (1–4445)	17.65 (24.30)	8.68 (0.03–118.8)	2.39 (1.22)	1.95 (0.24–6.29)
Replication	92.91 (113.41)	44 (3–618)	7.26 (6.58)	4.89 (0.33–38.63)	2.00 (0.97)	1.88 (0.24–4.36)

Note. IF = impact factor. [a]In cases where two initial studies were replicated by one replication study, the citation count of both initial studies was recorded.

which aligned very closely with psychology (0.27) and was a little higher than in education (0.16), which we calculated using data from Makel et al. (2012) for psychology and Makel and Plucker (2014) for education.

Although the total and annual citations were higher for initial studies, the citations of replications were far from low, despite this being a frequent concern about replication work. The mean annual citation of replication articles (7.26) was well above the mean impact factor of the journals publishing replications (2.00) and initial studies (2.39). It was also above even the highest journal impact factor in the Social Sciences Citation Index (SSCI) for linguistics (*Journal of Memory and Language*, 5.22) and education (*Educational Psychologist*, 5.69). This is compelling evidence that replications, at least those published to date, do not have low impact.[8]

We estimated that the mean number of citations of a study before its replication was published was 117.20, based on a mean of 6.64 years between an initial study and its replication, and a mean of 17.65 annual citations of an initial study. We acknowledge that this is an estimation based on an average evenly spread over time. For L2 research (where citation counts are generally much lower than, for example, psychology), we consider this to be a high number of citations before a study's reliability and generalizability are investigated via replication, especially given the large standard deviations in our data that indicate that some studies received many hundreds of citations before they were replicated.

In terms of the impact factor of journals that publish replications (Table 3), journals with both high and low 5-year impact factors published replications with no discernible association between impact factor and number of replications published, $r_s(26) = .157$, $p = .443$. On average, replications were published in journals with slightly lower impact factors than those of the initial studies, though with a small effect size whose lower 95% CI almost reached 0, $t(128) = 2.059$, $p = .042$, $d = 0.36$, 95% CI [0.01, 0.70]. This small difference would partly be due to the fact that just over a third (38.8%) of replications were published in the same journal as the initial study (compared with 30.6% in education and 19% in psychology).

Have Closely Interconnected Series of Initial and Replication Studies Been Conducted?

Our search identified 67 replications based on 70 initial studies. The mismatch in these numbers reveals some interconnectedness between groups of studies, where four studies replicated more than one initial study: DeKeyser and Sokalski (1996) replicated VanPatten and Cadierno (1993a, 1993b), Liu (1985) replicated Au (1983, 1984), Walters (2012) replicated Fitzpatrick and Meara

(2004) and Fitzpatrick and Clenton (2010), and Ellis et al. (2014) replicated both Ellis and Sagarra (2010) and Ellis and Sagarra (2011, Experiment 1). In these cases, the replications were conceived of (both by the authors and by us) as one replication. Further interconnectedness was found in two lines of research. First, Ellis and Sagarra (2011) served both as an initial study for the Ellis et al. (2014) replication and was itself a replication of Ellis and Sagarra (2010), and thus it was coded as both a replication and an initial study; and second, VanPatten and Cadierno (1993a) served as an initial study for DeKeyser and Sokalski (1996) and for VanPatten and Oikkenon (1996), and so it was coded twice in its capacity as an initial study. Overall, though, the interconnectedness of groups of studies was minimal, given that from the 67 self-labeled replications, only four were associated with more than one initial study, only one continuing line of replications was identified, and only one study was replicated more than once.

It is of course highly likely that more interconnectedness in L2 research exists than was evidenced in our search, due to unwillingness of researchers to self-label their studies as replications. Indeed, several of the initial studies were closely related to each other (close enough to be replicated simultaneously by one study) but did not self-label as replications themselves. However, it remains worrying that our sample only provided two clusters of studies that self-labeled as overt sequences of an agenda that extended beyond two studies (the VanPatten–Cadierno–DeKeyser cluster and the Ellis–Sagarra et al. cluster). Among other concerns, it suggests that the many syntheses and meta-analyses in the field (e.g., Plonsky & Brown, 2015, examined 81 meta-analyses) are bringing together studies that did not self-identify as replications of any kind. Meta-analysts seem to have observed this issue frequently because they have commented on the less-than-ideal comparability between studies in the domain under investigation (due to inconsistency of materials, measures, etc.) and it is one cause of the low number of studies in meta-analyses (e.g., Oswald & Plonsky, 2010, found a median of 16 studies reviewed in 27 meta-analyses in L2 research).

What Kinds of Studies Have Been Replicated?

Have the Findings From Initial Studies Tended to Be Statistically Significant or Null?

First, we checked the nature of analyses reported in the initial studies and found that, as expected, statistical procedures largely reflected NHST (mainly analyses of variance and *t* tests) that are normally used in L2 research (Plonsky, 2013; for details, see Table S3-2 in Appendix S3 in the Supporting Information

online). Next, we coded how the initial studies' findings were reported by the authors into four categories:

- null hypothesis rejected, which was usually reported as a finding of statistically significant difference/association between the variables under investigation with an alpha of .05;
- failure to reject the null hypothesis, which was usually reported as no statistically significant difference/association between the variables under investigation;
- trend/borderline differences/associations, as interpreted by the authors; or
- other, which usually indicated that statistical significance was not applicable to the research design.

This coding was necessarily broad brush, but the overwhelming finding was that researchers have replicated studies that had a statistically significant finding (87%), with only 3% of studies replicating a study with null results, 3% with a trend toward an effect, and 7% other. This suggests an influence of publication bias and/or the file drawer problem, even though we included initial studies that were not published in journals, on the assumption that books are perhaps perceived as being less prone to publication bias. Our finding is also possibly a consequence of (perceived or real) difficulties in interpreting null findings without ascribing methodological flaws to the study, which probably decreases the impetus to replicate studies with null findings.

In our view, these data fuel compelling arguments (a) to investigate the extent of publication bias generally by increasing overall replication effort (among other approaches); (b) to increase all types of replication (exact, partial, and conceptual) of studies with null findings, in order to inform theory and ascertain the extent to which initial null findings were indeed due to methodological flaws; and (c) to undertake peer review prior to data collection to reduce publication bias.

What Were the Designs and Contexts of the Initial Studies?

We examined characteristics of the initial studies to explore whether particular design features seemed to have a propensity to be replicated. The majority of replicated studies were one-shot, cross-sectional designs. However, more complex designs were also replicated, such as longitudinal (40%) and intervention (37%) studies.[9] In terms of context, 50% were laboratory based and 39% had collected data in a classroom (Table 4).

In terms of the measures used in the studies, the majority examined morphosyntax and used measures that were linguistic, written, and administered

Table 4 Study types/contexts in initial studies ($k = 70$)

Type/context	Initial studies
Laboratory	50%
Experimental/manipulated classroom	20%
Intact/ecologically valid classroom	13%
Lab plus intact or experimental class	6%
Not reported	9%
n/a	3%

Table 5 Measure and instrument types used in the initial studies

Feature	Initial studies[a]	Focus/type	Initial studies[b]	Modality/ mode	Initial studies
Morphosyntax	40%	Linguistic	87%	Oral	26%
Lexicon[c]	23%	Nonlinguistic	9%	Written	49%
Pragmatics	10%	Both	4%	Both	17%
Speech[d]	9%	Offline	83%	n/a	9%
Multiple features	19%	Online	9%	Comprehension	23%
Not reported, n/a	7%	Both	0%	Production	23%
		n/a	9%	Both	44%
				n/a	10%

[a]Adds up to more than 100% as some studies had more than one. [b]Throughout the article, unless otherwise stated, where a column (or row where applicable) does not add up to 100%, this is due to rounding error. [c]Including collocation and figurative language. [d]Including phonology, prosody, pronunciation, and fluency.

offline. However, overall, a very wide range of linguistic forms and assessments appeared in the initial studies (Table 5).[10] This variation in design characteristics and the finding that 67% of studies included a production measure and 43% had oral measures (which are usually more difficult to administer and/or score) suggest that L2 replication efforts have not tended to replicate only easier studies. Interestingly, although one might think that highly controlled, laboratory-based research would be more conducive to replication, studies with an online measure, such as self-paced reading or eye tracking, were rarely replicated in our sample ($k = 6$). This may reflect the relatively recent adoption of such techniques in mainstream L2 research (as found by Marsden et al., in press) but also the challenges posed by accessing and using expensive

hardware and software that is also comparable across sites and studies (as noted by Laws, 2016, experienced by Morgan-Short et al., 2018, and discussed, with practical advice, by Schmid et al., 2015). Infrastructure for collection of data via the Internet, such as that proposed by MacWhinney (2017), would help to alleviate this problem.

What Were the Participant Characteristics in the Initial Studies?

Participant characteristics, such as age, language background, and proficiency, also provide critical insight into the kinds of studies that tend to be replicated. In terms of language proficiency, we found that of the 62 initial studies with language learners, 17 gave some indication of whether participants were beginner, intermediate, advanced, or a combination of these.[11] However, 25 did not specify the proficiency level, and 20 studies were coded "other" for a range of reasons (e.g., the study gave number of years of learning experience, rather than proficiency). In terms of ages, 29 studies used university students without specifying ages, which in reality vary enormously but typically range between 18 and 30 years. Of the 22 studies that did report participants' age, we calculated a mean of 22.18 years ($SD = 11.68$).[12] Finally, most initial studies involved English as the target language. There was a little more variation seen in participants' L1, though seven studies did not report the participants' L1 (see Table S3-3 in Appendix S3 in the Supporting Information online). In all, replications have been largely of initial studies with young adult learners of English, in line with previous observations about participant demographics in L2 research (Ortega, 2013). Most critically for the current study, our data (or lack thereof) clearly demonstrate how unclear reporting practices have adverse consequences for replicability because replicators cannot know what sample population to target, which characteristics they may wish to intentionally change, or which characteristics they should acknowledge as being different from the initial study.

To What Extent Were the Materials of the Initial Study Accessible?

A final feature related to the kind of studies that have been replicated is the degree to which initial studies are transparent in terms of materials. We found that 17% of initial studies did not provide any materials at all and that 41% provided only partial examples in the text of the article. Although 37% did provide at least one full instrument, these did not provide all of the instruments used to collect the data that were ultimately analyzed in the study. Only three of the studies in our sample provided a full set of materials (Table 6).

Table 6 Availability of materials in initial studies

Material availability	Initial studies	Behind journal paywall[a]	Open access	n/a or other
No materials	12	–	–	–
Partial examples	29	90%	3%	7%
One full instrument (not all materials)	26	73%	8%	19%
Full materials used for analysis	2	100%	0%	0%
All full materials	1	100%	0%	0%

[a]When materials are available behind a journal paywall, this does not make replication easy as not everyone has access to all journals (e.g., researchers in certain socioeconomic contexts or practitioners without journal subscriptions). Additionally, it is possible to acquire some articles via open access portals, and so know about a study but not have access to its materials, which can remain behind journal paywalls in supplementary materials.

Table 7 How materials were made available to the replicators

Availability in initial study (*k* replications)	In article	Passed on in private[a]	Shared authorship[b]	Unclear
No materials (12)	0%	25%	33%	42%
Partial examples of an instrument (26)	54%	12%	31%	4%
One full instrument (26)[c]	85%	4%	19%	0%
Full materials used in analysis (2)	50%	50%	0%	0%
All full materials used in entire study (1)	100%	0%	0%	0%
Total (67)	54%	12%	25%	9%

[a]Acknowledgment sections were searched to determine whether researchers were thanked for materials. [b]Materials were not available with the initial article or open access, so we assumed materials were passed on via the author(s) common to the initial and replication studies. [c]Adds up to more than 100% because two studies had an instrument in the article and had shared authorship.

Our data regarding the availability of materials beg the question of how replicating researchers acquired the materials needed to replicate the study. In our sample of replication and initial studies, it was often unclear how materials had been obtained or whether they had been recreated, especially in cases where no materials or just examples were available (Tables 6 and 7). Thus, replication studies seemed to have been carried out even when materials were not available or were only described. As with gaps in reporting about participant

Table 8 Replications with changes to participant demographics ($k = 67$)

	Participant characteristic			
Type of change	L1	L2	Proficiency	Age
No change	43%	76%	39%	58%
Claimed constant, but coder identified change	0%	0%	0%	0%
Change motivation for replication	31%	13%	15%	1%
Change acknowledged, not motivation for replication	9%	7%	10%	3%
Change not acknowledged	9%	3%	6%	19%
Unclear/not reported or n/a	7%	0%	30%	18%

characteristics, poor availability of materials reduces the replicability of studies and also weakens claims that can be made by replications (because the extent of parity with initial studies is difficult to ascertain).

To What Extent and How Did Researchers Change the Initial Studies?

What Are the Overall Extent and Types of the Changes?

In the narrative review, we noted that a limited number of motivated changes between an initial study and its replication, such as those often suggested as future directions by the initial study authors, can be desirable for systematic research agendas but that too many changes or changes that are unmotivated or unacknowledged impede the ability to account for differences in the findings between studies. To gain insight into the types and numbers of changes between initial and replication studies, we coded and counted each change between pairs of studies. We distinguished among three types of changes: (a) changes that were overtly reported as intentional alterations that explicitly motivated the replication, as one would expect in partial and conceptual replications, henceforth referred to as *motivated changes* or a motivation for replication; (b) changes that were acknowledged by the authors but not explicitly articulated as principled motivations for the replication, henceforth referred to as *acknowledged changes* but not motivations for the replication; and (c) changes that were noted by our coders but not acknowledged by the authors, henceforth referred to as *unacknowledged changes*.

In terms of changes to participant characteristics (Table 8), the participants' L1 was the most common, often as an intentional change motivating the replication ($k = 21$) or an acknowledged but unmotivated change ($k = 6$). There were a few instances of motivated changes to participants' L2 or level of L2 proficiency. Reassuringly, there were no instances where authors overtly claimed

that participant characteristics were constant between studies but where the coder thought there had been a change. However, there were several instances of unmotivated or unacknowledged changes. For example, 6 studies changed the L1, 6 the proficiency, and 19 the ages of the participants without explicitly acknowledging these differences.

For linguistic features, mode (production/comprehension), and modality (written/oral), we observed surprisingly few changes, with only about one in five of the replications amending one or more of these characteristics (see Table S3-4 in Appendix S3 in the Supporting Information online). However, about half of the replications changed the outcome measures in various ways, such as using different items, tasks, stimuli, or proficiency measures or manipulating whether a test was done in pairs or in a group. A quarter of studies made such changes to the measures that were either not motivated or not acknowledged. Changes to measures were often justified as improvements to the data-elicitation techniques used in the initial study. Thus, one reason might have been poor instrument or coder reliability found in the initial studies. However, indices of reliability, such as Cronbach's alpha, percent agreement, or Cohen's kappa coefficients, were reported in only 17% ($k = 12$) of the initial studies. Thus, changes to instruments appeared to be largely based on the replicating researchers' subjective evaluation of the instruments.

The extent and purpose of these changes cause concern. For example, changing the data-elicitation instrument is a significant change, best conceived of as an intentional alteration that motivated a replication. Such changes can, if they are not an intentional design feature (which was the case in a quarter of our replication studies), constitute a major threat to interpretability, particularly in cases where findings are different between the studies. Of course, there is a tension between changing a measure for perceived improved internal validity and compromising the initial study's characteristics and, therefore, the capacity to determine the cause of differences in findings. To us, these findings underscore the need to continue refining and sharing the field's measurement toolkit to reduce the need to change measures between interconnected studies and thus to increase parity between these studies. Indeed, this goal was one of the main purposes behind establishing the IRIS database of research materials (Marsden et al., 2016).

In terms of study design more generally, we observed very few changes (see Table S3-5 in Appendix S3 in the Supporting Information online). However, 23% of replications made changes to the study's context, that is, a L2 versus a foreign language context (though this change was motivated for only 10% of replication studies). Researchers largely maintained the longitudinal or

cross-sectional designs of the initial studies, with just three exceptions. There were, again, instances where changes were not acknowledged, the most concerning of these being in the domain of the statistical analyses, with over a third of studies using different statistical procedures without clearly justifying this change. Although some of these changes were appropriate, given the other changes made by the replication, the explicit acknowledgement to the reader was inconsistent.

Another change that may occur between initial and replication studies involves the sample size. We found that the subgroup sample size of replications was a mean of 4.4 ($Mdn = 1.4$) smaller than that of the initial studies, with a very large standard deviation (51.4) and a wide range from –304.0 to 108.5.[13] As noted earlier, smaller sample sizes in replication studies compared to the initial studies can be problematic if effects observed in the initial study are not observed in the replication because this difference in effects could be accounted for both by lower power and/or a genuinely different finding. Despite this concern and variation in sampling practices, sample sizes in replication research generally seemed to be higher than the averages found in other, broader syntheses of L2 research: mean study sample size of 114.4 for initial studies and 88.1 for replications and mean subgroup sample size of 41.1 for initial studies and 36.4 for replications. These results compare favorably with those obtained by other researchers: the subsample median[14] of 19 reported by Plonsky (2013), the subsample mean of 22 reported by Plonsky and Gass (2011), the medians per condition of 26 (within-subject designs) and 20 (between-subject and mixed designs) reported by Lindstromberg (2016).

Collapsing across the types of changes (Table 9, last row), there was, per replication, a mode of (a) two motivated changes, (b) one acknowledged but not motivated change, and (c) two changes that were not acknowledged by the authors. Overall, our findings suggested that in much L2 replication work to date, there have been about as many or more unmotivated and unacknowledged changes per study as motivated changes. As such, it would currently be difficult to make any general evaluation of the reproducibility of L2 research.

To What Extent Did the Amount of Change Between Initial and Replication Studies Relate to Nomenclature of Replications?

Given that there was such variability in nomenclature (Table 1) and that the majority of studies are simply self-labeled as replication with no further qualification, we were unable to statistically examine the numbers of changes as a function of the sublabels of replication. Descriptively, we were not able to find any clear discernible patterns. For example, in the three studies that called

Table 9 Replications making different types of changes to the initial studies

	Number of changes								
	0	1	2	3	4	5	*M*[a]	*SD*[a]	*Mode*[a]
Claimed constant, but coder identified change	94%	4%	1%	0%	0%	0%	0.07	0.32	0
Change motivation for replication	33%	28%	21%	9%	7%	1%	1.34	1.31	2
Change acknowledged, not motivation for replication	45%	31%	13%	9%	1%	0%	0.91	1.04	1
Change not acknowledged	46%	25%	15%	9%	4%	0%	1	1.18	2

Note. [a]Number of changes per study.

Table 10 Extent to which replications supported the findings of the initial studies ($k = 67$)

Level of support	Replication
Not supported	15%
Partially not supported	13%
Partially supported	34%
Very supported	34%
Unclear	3%

themselves conceptual replications, where one could expect several and all types of change, we found very different patterns. Specifically, two conceptual replications had no motivated changes whereas the other had three; regarding changes acknowledged but not a motivation for the study, one had none, one had one, and the other had three; and, finally, regarding unacknowledged changes, two conceptual replications had two and one had four. Our one self-labeled close replication (Waring, 1997) perhaps fit the expected profile, having one change that motivated the replication and no other changes to key variables. Table S3-6 in Appendix S3 in the Supporting Information online includes the two sets of self-labels that had the largest number of groups in our sample: *partial* ($k = 14$) and *replicat** without a qualifier ($k = 45$). The data show that the amount of change seems to be similar regardless of the label. We acknowledge that the low number of partial replication studies precludes firm conclusions, but at the very least the data demonstrate little systematicity of nomenclature. This replication self-identity crisis is arguably one cause of the lack of self-labeled replication studies published in the field, as authors, reviewers, and editors vary in their understanding of what does and does not constitute (different types of) replication.

To What Extent Did Replications Support the Findings of the Initial Studies?

To examine this question, we used a 4-point scale to code the extent to which the initial study's findings were supported by the replication as claimed by the authors of the replication (Table 10):[15]

- 0 = not supported (results did not support the initial findings at all),
- 1 = partially not supported (the majority of results did not support the initial findings),

- 2 = partially supported (the majority of the results supported the initial findings), and
- 3 = very supported (results supported the initial findings).

We found that most studies (68%) presented findings that generally supported the initial studies, a finding that aligns closely with the findings of 67.4% for education (Makel & Plucker, 2014) and, more loosely, to the 78.9% found for psychology (Makel et al., 2012). That is, just under a third of our replication studies produced findings that were divergent from the initial study, arguably demonstrating the basic need for replication research to corroborate the validity of findings in L2 research generally. However, supportive or nonsupportive findings from studies that were not direct replications (as in the current synthesis) cannot provide a meaningful indication of reproducibility in the field because many of the replication studies introduced substantial heterogeneity into their design, either intentionally or not.

How Did Authors Compare Their Findings With the Initial Findings?

We explored how replicating authors compared their findings with the initial study's findings by coding for two main issues. First, we coded how the initial study's data were presented by the replicators (Table 11) and found that only about a quarter presented descriptive statistics from the initial study and that even fewer used other types of statistics or data from the initial study. Whatever this is due to (e.g., space constraints, an assumption that reviewers and readers will access the initial article, or lack of incentive to report fully), it renders basic comparisons between studies difficult.

Second, we coded for how the data from both studies were compared (Table 12) and found that comparisons between the studies were generally narrative or based on a dichotomous interpretation of NHST, for example,

Table 11 How replications presented and used the results from the initial study ($k = 67$)

Provided descriptive statistics	28%
Provided inferential statistics	13%
Extracted reported data, analyzed with replication data	12%
Provided effect size	6%
Used raw data in a new statistical analysis	6%

Table 12 How replications drew comparisons with the initial studies ($k = 67$)

Narrative comparison	93%
Mentioned findings of initial study	90%
Based on dichotomous interpretation from NHST	84%
Compared descriptive statistics	34%
Unclear	6%
Compared effect sizes	1%

Note. NHST = null hypothesis significance testing.

findings were significant or not. These two observations—that comparisons were almost exclusively narrative or based on NHST and that so few analyses used the initial study's data—are hardly surprising given the lack of availability of effect sizes and raw data in the initial studies.

Effect sizes, as authors have noted many times (e.g., Norris et al., 2015), are useful because they enable comparisons to be made using standardized units across studies to interpret the magnitude of difference or association in meaningful paired comparisons. Morgan-Short et al. (2018) provided an example of a study giving independent effect sizes for intersite comparisons and aggregated effect sizes in an intrastudy meta-analysis of direct replications (see also Ellis & Sagarra, 2011). In our sample of 70 initial studies, Cohen's *d* was provided in seven studies and *r* by one study, whereas 81% did not provide any effect size values.[16] We also did not find instances of replicators extracting effect sizes from the initial studies (e.g., Cohen's *d* can be calculated from *t* and *F* statistics when two groups are compared). We found it surprising that the use of effect sizes had not become more embedded by the time of this review, given that many of the initial and most of the replications happened after the influential meta-analysis by Norris and Ortega (2000) emphasizing the importance of effect sizes and after several journals started requiring the provision of effect sizes.

As noted above, there were small numbers of studies that used the raw data from the initial study in the replication's analysis. Such access to data was possible because all four studies had author overlap (Table 11). Interestingly, three of these found very supportive evidence for the initial study. The fourth study, Ellis and Sagarra (2011), found evidence partially not supporting Ellis and Sagarra (2010) and was the only study to use Cohen's *d* to draw comparisons. This brings us to the question of the factors—including that of author overlap—that may be associated with replication studies producing findings that supported the initial findings.

Table 13 Replications that are supportive/not supportive of initial findings, as a function of author overlap

Author overlap (k studies)	Not supportive	Partially not supportive	Partially supportive	Very supportive	Not reported/ clear
No overlap (46)	20%	17%	30%	28%	4%
Some overlap (21)	5%	5%	43%	48%	0%
Total replications (67)	15%	13%	34%	34%	3%

Author Overlap

We first quantified the amount of author overlap in our sample of studies and found that 6% ($k = 4$) of the replications had the same authorship as the initial study, 25% ($k = 17$) had some authorship overlap (one or more authors in both the initial and replication studies), and 69% ($k = 46$) were carried out by entirely new author teams.[17] This could imply a degree of independence in the replication research in our sample. We explored various effects that overlap in authorship may have had on the extent to which replications supported the initial findings.

First, as seen in Table 13, authorship overlap seemed to be associated with supportive findings. When there was no author overlap between the initial and replication studies, 37% ($k = 17$) of replication studies were generally not supportive and 59% ($k = 27$) were generally supportive whereas with some author overlap, only approximately 10% ($k = 2$) tended not to be supportive and 90% ($k = 98$) were supportive. This pattern was statistically significant, $\chi^2(1) = 5.824$, $p = .016$; likelihood ratio $= 6.634$, $p < .01$. It also aligns with the ratios found by Makel et al. (2012) in psychology (91.7% supportive with author overlap, 64.6% supportive without) and Makel and Plucker (2014) in education (88.7% with, 54% without).

There are several explanations for these data for author overlap. They could reflect questionable research practices, which may be more likely if an initial study author is biased toward finding a particular outcome in the replication. They could (also) be a consequence of greater fidelity to the initial study because materials were available and protocols were more strictly adhered to. This might be because author overlap could incur fewer researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011), that is, a reduced likelihood of divergence at the many decision points in any study. There may (also) be a possibility that replications with author overlap might be more likely to have a confirmatory aim (and therefore be closer to the initial study, with

Table 14 Studies with total number of changes, as a function of author overlap

	Changes were motivation for replication			Changes acknowledged, not motivation for replication		Changes not acknowledged	
Number of changes	0	1	2+	0	1+	0	1+
No overlap ($k = 46$)	30%	28%	42%	41%	59%	39%	61%
Some overlap ($k = 21$)	38%	29%	34%	52%	48%	62%	39%
Based on k replications	22	19	26	30	37	31	36

fewer changes), rather than the aim of testing generalizability by intentionally manipulating several variables.

To further investigate this last possibility, we compared the number and type of changes between replication and initial studies as a function of author overlap. Although author overlap did not seem to be associated with the proportion of studies that changed just one feature as a specific motivation for the replication, we found that, overall, replications with author overlap tended to make fewer changes to the initial studies (Table 14). First, there were slightly more replications with author overlap than without overlap that made no changes of any type (motivated, acknowledged, and unacknowledged). Second, there were more studies without author overlap than with overlap that made several unmotivated or unacknowledged changes. This indicates closer replications (i.e., involving less heterogeneity) with author overlap, which (intuitively at least) seem more likely to produce findings that are more in line with the initial studies. Thus, the extent to which replications supported the initial findings could, at least partially, be accounted for by the trend that replications with author overlap were closer to the initial studies than those without author overlap.[18]

Amount of Change From the Initial Study

It may be that increased heterogeneity—quantified as the number of changes between the replication and initial studies and independent of authorship overlap—could be linked to a lower likelihood of replications producing findings that supported the initial studies. However, the data shown in Table S3-8 in Appendix S3 in the Supporting Information online suggest no strong or interpretable patterns in this matter. This broadly aligns with the lack of evidence in psychology and education research that direct replications were any

Table 15 Supportiveness of replications as a function of the availability of the data collection instruments

Replication result	None	Examples	One full
Partially not supportive ($k = 18$)	42%	37%	16%
Partially/very supportive ($k = 44$)	9%	41%	48%
Based on *k* replications	12	26	24

more (or less) likely to support initial findings than conceptual replications (Makel et al., 2012; Makel & Plucker, 2014). It also chimes with the negligible to small effects of heterogeneity found in the Many Labs project (Klein et al., 2014). These findings suggest that other issues may be more strongly linked to the extent of supportiveness, such as the nature of the effect under investigation (as argued by Klein et al., 2014), the theorized intention of the heterogeneity or, perhaps, as examined below, the transparency of the initial study's materials.

Transparency of the Initial Study's Materials

Finally, it may be that without access to full materials from the initial study, replicating researchers need to create their own materials. This would introduce unintentional and unacknowledged heterogeneity between studies, which could in turn account for less supportive findings. Thus, we examined whether the availability of the initial study's materials was associated with supporting its findings (Table 15). Of the 65 studies that could be included in such an analysis, we observed that instrument transparency was associated with an increased likelihood of replications producing supportive findings, a pattern that was statistically significant, $\chi^2(1) = 11.489$, $p = .003$.[19] We think that this provides some evidence for one benefit of making materials transparent. Overall, regarding the factors that are associated with reproducibility, our results seem to suggest that author overlap and the availability of materials were associated with supportive findings whereas the number of changes between initial and replication studies was not.

Further Discussion and Recommendations

In light of these narrative and systematic reviews and our own experiences with replication work, we summarize key findings and propose a set of recommendations. Our discussion and recommendations align with the four main themes addressed by both the narrative and systematic reviews above, though these themes are fragmented into seven subsections here and presented in a

slightly different order: (a) the quantity and nomenclature of replication (Recommendations 1, 2, 3, and 4); (b) changes between initial and replication studies (Recommendations 4, 5, and 7); (c) the warranting of what research gets replicated (Recommendations 6, 7, and 10); and (d) the extent of reproducibility and its relations with author overlap, materials transparency, and heterogeneity between replication and initial studies (Recommendations 7, 8, 9, 10, and 11). In line with our aim to consider infrastructural challenges to replication research, Recommendations 11 to 16, along with Recommendations 2 and 9, allude to infrastructural and cultural needs in publishing, funding, and training. All recommendations are united by the aims of increasing the quantity and improving the quality of replication research in the field of L2 and multilingualism research.

Increasing the Amount and Speed of Replication

Although we cannot determine an optimum rate of replication or an ideal balance between replication and innovation, our data certainly demonstrate an extremely low rate: Replications have constituted approximately 1 out of 400 articles in those journals that have published at least one self-labeled replication in L2 research since the first published L2 replication in 1973. Critically, this rate would be much lower if it could be calculated using the whole, larger set of journals that ever publish L2 research and from the start of their history. Makel et al. (2012) and Makel and Plucker (2014) were able to calculate this broader denominator easily and objectively by using the set of journals delineated by the discipline categories of education and psychology in the ISI Web of Knowledge Journal Citation Reports whereas there is no such discipline-specific list for L2 and multilingual journals. Even more worryingly, despite our more generous calculation, the rate we found was much lower than that in psychology, the key parent discipline for L2 research that adopts quantitative, hypothesis-driven approaches and a discipline that is itself concerned that its own replication rate is too low. Our data also demonstrate a slow speed of replication. The observed mean gap of 6.4 years is not likely to expedite the checking and refining of theories before implications for academic and practitioner communities take root. As argued by Makel and Plucker, "science may be self-correcting, but the often glacial pace of that correction does not match the speed of dissemination when results enter the public consciousness" (p. 313). We are unequivocal in our first and overarching recommendation.

> Recommendation 1:
> Increase the number of replication studies and the rate at which they are performed and published.

We also emphasize that data are needed about the causes of low published replication rates to inform our efforts, including those recommended in the following sections, in empirically grounded ways. For example, the publication of replication studies that had null findings or that did not support the initial findings may have been adversely affected by publication bias and so may be one cause of the overall low rate of published replications. Initiatives such as Positively Negative (PLOS, 2015), an open collection of studies with null or inconclusive findings, which includes studies labeled *failure to replicate*, may be useful and worth evaluating. There are many other potential causes of the low rate of replication research, such as low prestige and a related unwillingness to self-label as a replication.

> Recommendation 2:
> Make systematic inquiry into the causes of low rates of published replication studies and provide (more) empirical evidence about the extent and causes of publication bias in the field.

The Importance of Nomenclature

The low rate of replication is likely due in part to a lack of willingness to self-label as replication (Neulip & Crandall, 1993; Polio, 2012b). This reticence is complex. Anecdotally, we observed during colloquia discussing this study and the research by Morgan-Short et al. (2018) that some researchers reported actively undertaking and promoting replication with students and in their own work, yet they were less enthusiastic about labeling these studies as replications. Here we illustrate with three relatively recent examples of what we think is fairly standard practice. This is, we stress, not to criticize these studies (and a good proportion of our own research certainly has aligned with this practice). Rather, we aim (a) to acknowledge that our synthesis is not a fully comprehensive reflection of the amount and nature of replication effort in the field and (b) to recognize the complexities that our arguments and recommendations about nomenclature entail. First, Kim and Nam (2017) had closely related aims and used the same tests and similar analyses procedures as Ellis (2005). They did not self-label as a replication (their title used the term *revisited*) and yet referred to three other studies that used the same materials in the same agenda as replications though none of those studies was self-labeled a replication. Second, Trenkic, Mirkovic, and Altmann (2014) did not self-label their study as a replication, but their aims, design, and stimuli were closely informed by the study by Chambers, Tanenhaus, Eberhard, Filip, and Carlson (2002), who were acknowledged for sharing stimuli, and they reported that their findings replicated the findings of Chambers et al. Third, several large,

coordinated studies have used the same (or very similar) shared materials across different sites (e.g., Bergmann, Meulman, Stowe, Sprenger, & Schmid, 2015; Dimroth, Rast, Starren, & Watorek, 2013; Meulman, Wieling, Sprenger, Stowe, & Schmid, 2015; Schmid, 2011). None of these studies was retrieved by our search, even though they are likely examples of partial or conceptual replications because they sought to make claims about replicating previous aims and findings and used or adapted materials from earlier studies. Thus, arguably, our estimate of the amount of published replication research in the field underrepresents the total wider replication effort when it is more broadly defined.

Frank debate is required about the advantages of three broad approaches to nomenclature: (a) using an agreed system of replication labels in titles or abstracts; (b) maintaining a looser range of other terms for indicating replication efforts, such as *exten** (McManus & Marsden, 2017; Nakamura, 2012) or *revisit** (Au, 1983; Kanno, 2000); and (c) alluding to closely related theoretical and methodological precedents (often covertly) within study reports. Here, we present arguments that a reticence to label with the term replication is detrimental for the field. First, it hinders the general tracking of intellectual connections and hides theoretical and methodological precedents under an invisibility cloak or a "cloaking device" (Makel et al., 2012, p. 541). Second, without replication labels, heterogeneity from one study to the next can pass largely unchecked. We found that despite many suggestions in the limitations/further research sections of articles regarding necessary replications with different language combinations, participant demographics, or design features, very little such specific variation is undertaken systematically in self-labeled replications, and variation was often accompanied by other, potentially confounding, changes. In contrast, using the label replication establishes a need for both the researchers and reviewers to monitor interstudy variation and identify precise relationships with one or more specific study/studies. This in turn increases the field's ability to confirm and reject theories across studies. Finally, the lack of self-labeling adversely affects efforts to synthesize and meta-analyze research. Among other purposes of synthesis, better self-labeling would facilitate future efforts to examine reproducibility in the field of L2 acquisition to ascertain the reliability and generalizability of findings (as has been done by the large-scale replication efforts in psychology).

In sum, we argue that explicit identification, via self-labeling with some replication nomenclature in titles or abstracts, clarifies the relationships between studies, and this would help the quality and scope of research agendas. For example, it would (a) render theoretical and methodological precedents

more visible; (b) facilitate reviewers' evaluation of the extent of changes to previous research procedures and materials; (c) encourage more tightly knit series of interconnected studies by requiring researchers to explicitly operationalize and articulate changes to earlier research; and (d) improve the quality of syntheses and meta-analyses by, for example, facilitating the comparability of studies.

Recommendation 3:
Use more self-labeling with the term replication wherever appropriate.

In terms of subtypes of replication, we found a very wide range of labels and negligible relations between these labels and the amount or type of change between the initial and replication studies. We thus propose a simple distinction based on the principle that direct replications aim to test data and analysis (i.e., to confirm previous findings via a study with, as far as possible, the same conditions) whereas partial replications test a construct by manipulating one of the initial conditions or study characteristics to test generalizability to one new context/condition (see Lykken, 1968). Thus, direct replication would describe a study in which there was no intention to change any variables deemed likely to affect results (according to current knowledge). Because minor deviations from the initial study can be unavoidable, especially with human participants, any such heterogeneity would be reported as fully as possible.

Partial replication, on the other hand, would describe a study that intentionally changes only one significant component of the initial study to check a priori for one well-defined boundary condition or moderator of the initial findings. This could include a principled change in instrumentation, analysis, linguistic form, or a participant characteristic. In our study sample, partial replication was the most frequently used sublabel. And although we are confident that the term is already in our nomenclature, we recommend that its usage/function become more consistent. Conceptual replications introduce more than one significant change to the initial study and can extend agendas in multifaceted ways but are in a weaker position for ascribing different findings to the adaptations made to the initial study. However, retaining this label is, we think, helpful for authors and reviewers seeking to identify the extent of relations between studies.

Recommendation 4:
Apply a principled, standard nomenclature as follows: *Direct replications* make no intentional change to the initial study and seek to confirm methods, data, and analysis; *partial replications* introduce one principled change to

a key variable in the initial study to test generalizability in a clearly predefined way; and *conceptual replications* introduce more than one change to one or more significant variables. In all cases, ensure that potential heterogeneity and contextual details are documented as fully as possible.

Ascertaining the extent and nature of change between initial and replication studies can be severely hampered by unclear reporting. For example, we found that some of the changes that replications introduced were not acknowledged by the authors. We recommend that authors of replications clarify the relationship with the initial study (including descriptive statistics and effect sizes) and combine analyses where possible. However, even with better reporting of the methods, data, and findings of the initial studies, it is unlikely that the replication articles can do full justice to the initial report, especially given that some journals assign replication studies to a shorter article type. In view of this, reviewers need to be familiar with the initial study and read it alongside the replication to be able to corroborate the claimed relationships. This will have implications for authorship blinding practices in cases where there is author overlap between the initial and replication studies.

Recommendation 5:
Reviewers of replications should also read the initial study that is being replicated.

Warranting What Should Be Subject to a Replication Study

Various propositions exist to set benchmarks or define rationales for when a study merits replication, such as a citation metric or the co-occurrence of specific characteristics, for example, low sample sizes, large effect sizes, marginal/borderline statistical significance, or unexpected findings (Lindsay, 2015). However, we do not propose a set of such benchmarks because these may become overinterpreted (as have been an alpha level of .05 and small/medium/large effect sizes) and could exacerbate the image of replication as an unoriginal, mechanistic undertaking. Part of the skill in replication work is surely choosing studies worth replicating and justifying this to reviewers and editors. These justifications are likely to include citation counts; low sample size; surprising results; and theoretical, methodological, or practical issues, but a rigid formula based on a fixed composite of these is likely to be cumbersome and unreliable. Thus, we suggest that there should be little or no top-down (e.g., journal or professional association) control, and researchers' agendas should drive what is replicated.

> Recommendation 6:
> Provide warrants for replication studies and have them peer reviewed on a case-by-case basis with rationales including, but not restricted to, one or more of the following characteristics of the initial study: surprising findings; one or more troubling methodological features; and/or high (potential) impact, such as theoretical or practical significance.

A related phenomenon that may, however, require top-down influence is the rate of published studies that replicate initial studies with null or borderline findings because the current synthesis found a paucity of such replications—just 4 of 67. We do not suggest that this should be addressed by a blanket recommendation, such as "increase attempts to replicate initially null findings," because the phenomenon is tightly related to the low rate of publication of studies with null findings in the first place, which is in turn influenced by publication bias. However, given that replication can increase the interpretability, and therefore the value, of initial null findings, we suggest that these issues are certainly worthy of empirical investigation (see Recommendation 2).

Collaborative Ethic to Sustain an Independent Replication Effort: Transparency of Materials and Data

Several issues will determine the extent and speed with which we can adopt more collaborative approaches to facilitate replication. We found that changes to stimuli, instruments, and measures (such as elicitation tests) were relatively frequent between an initial and a replication study. Although these changes were sometimes intentional, being a motivation for the replication, often this was not the case. This is a chief concern because measures often constitute the key dependent variables, and changes to them reduce comparability with previous research (Marsden et al., 2016; Thomas, 1994, 2006). For example, several meta-analyses have shown that effects of instruction vary as a function of measurement type (e.g., Lee et al., 2015; Norris & Ortega, 2000). Another problem we found was that the extent of change could not be ascertained due to omissions in the initial study's report and lack of availability of materials and data. Methodological transparency can improve these problems, facilitating replication and improving its quality and reliability (Marsden et al., 2016). Another motivation to make materials fully available is, arguably, that—according to our findings—the more that materials were available, the more likely a replication was to find support for the initial study. Transparency may also influence replication in other ways that require further investigation, such as increasing the quantity of replication due to ease of accessibility to materials. Further,

there is emerging evidence (Plonsky et al., 2017) that positive correlations exist between transparency of research materials (i.e., number of entries on IRIS) and journal citation counts. With high citation being one factor that can trigger a replication, it seems that transparency of materials could be associated with increased replication research. (We note that a range of factors may cause methodological transparency itself).

> Recommendation 7:
> Increase open availability of materials, including proficiency measures, for L2 research.

In addition, sharing data is essential for cumulative analyses that join data sets and examine moderator effects of interstudy variation, which is especially important given the well-documented lack of power in L2 research (e.g., Plonsky, 2013, 2015). We found only one bundle of self-labeled replications that conducted an internal meta-analysis, which was possible because the researchers used the same materials and had the data from the initial study fully available (Ellis et al., 2014; see also Lindsay, 2017; Morgan-Short et al., 2018). Making data available entails ethical considerations (e.g., institutional review boards) early in the research process and is not possible in all situations, but it is increasingly a requirement of funders.

> Recommendation 8:
> Make more research fully transparent and open for replication by making data available.

Researchers, reviewers, and editors all have the responsibility of improving our collaborative ethic. Trofimovich and Ellis (2015) adopted the Open Science Badges for *Language Learning*, and several other journals now also value open materials and data in this way (e.g., "Author guidelines for contributors," 2017; "Instructions for contributors," 2017). Kidwell et al. (2016) and Giofrè, Cumming, Fresc, Boedker, and Tressoldi (2017) have provided quantitative evidence for the effectiveness of this initiative for the transparency of materials and data. Indeed, partly as a result of these initiatives and push from journal editors, IRIS now holds 24 sets of L2 data, in addition to approximately 3,600 files of materials and analysis protocols. Although we have high expectations that transparency via materials and data sharing will improve the quality and quantity of replication efforts, there is still much work to be done in these endeavors. For example, Marsden et al. (in press) found only 4% of

self-paced reading studies had openly available materials, and 77% had only a brief example of stimuli available in their articles.

> Recommendation 9:
> Encourage more journals to give more and stronger incentives to their authors for systematically making materials and data openly available.

Independence Combined With Professional Practice and Collegiality: Authorship Practices

Our observation that supportive findings from a replication study were significantly more likely when authorship overlapped between the initial and replication studies, compared to independent replication, aligned very well with those of Makel et al. (2012) and Makel and Plucker (2014) from two related disciplines. We do not make conclusive claims about why author overlap tended to be linked to more supportive replications because this could be accounted for by increased questionable research practices and/or by reduced heterogeneity due to access and fidelity to materials and fewer researcher degrees of freedom. However, we argue that each of these explanations is concerning because reduced heterogeneity should be possible without overlapping authorship so that reproducibility of findings would be unrelated to author overlap. We suggest that replication carried out independently from the initial studies is desirable to reduce any influence that author overlap may have on our insight into the reproducibility of L2 research findings. Thus, when materials and data for initial studies are available, author overlap would become a matter of collegiality rather than necessity.

However, independent replications can be perceived negatively as bullying, as discussed by Bohannon (2014). Inviting the initial author to review replication studies can help reduce this, and (in the case of a Registered Report) the initial authors can be invited to provide a Stage 1 review before data collection (see Marsden, Morgan-Short, Trofimovich, & Ellis, 2018). Even more transparent practices that may promote more and higher quality replication, reduce publication bias, and reduce perceptions of bullying include (a) publishing open reviews and authors' responses to reviews (e.g., in *BMC Psychology*; Laws, 2016); (b) giving initial authors an automatic right to a peer-reviewed published commentary (e.g., in *Perspectives in Psychological Science*; in our sample, we found one such example, Kanno, 2000); and (c) adversarial collaborations (Coyne, 2016; Kahneman, 2014; Koole & Lakens, 2012; Mellers, Hertwig, & Kahneman, 2001), where researchers who account for phenomena differently agree to work together following a single protocol. We raise

awareness of the existence of these more extreme measures but hope that the other mechanisms that we recommend, such as transparent materials and data and the reviewing of methods prior to data collection, serve to reduce any perception of bullying that independent replication may engender.

> Recommendation 10:
> When possible, ensure that replication studies are conducted by researchers independently of the initial study's authors but that the initial authors are invited to be involved at some stage of the review process, preferably prior to data collection (see Recommendation 12 about Registered Reports).

For multisite replications, authorship practices may be required that are relatively rare to date in L2 research. The large multisite efforts have thus far been in fields where large authorship teams are the norm. In line with these practices, Morgan-Short et al. (2018) offered coauthorship to those collecting and entering data and running predefined analyses, with lead authorship for those convening the multisite replication, providing the protocols, and formally reporting the results. Even with this coauthorship agreement, they were fortunate in securing collaborators, and a reciprocal ethic is needed to support such large-scale multisite replication efforts (such as "I collect data for others; others collect data for me"). Formal infrastructure is likely to help here, such as the Call for Replication Collaborators button on IRIS and the Centre for Open Science's Study Swap (https://osf.io/view/StudySwap), whereby researchers seek collaborators or offer participant availability.

> Recommendation 11:
> Increase multisite collaborative replication efforts.

Cultural and Procedural Changes in Publishing

Various initiatives are available to increase the amount and quality of replication, the most obvious of which is perhaps author guidelines of journals explicitly encouraging replications. However, our data suggest that this alone was not a reliable or necessary mechanism. Although the journal which had published the most replications had a statement inviting replications, the other three journals with such a statement actually published fewer replications than journals without such a statement. Indeed, journals that simply state that they publish replications reach only Level 1 of the TOP Guidelines on replication (Nosek et al., 2015a).

Another mechanism might be the idea of an Accountable Replication Policy (proposed by Chambers, 2016, launched at *Royal Society Open Science* in January 2018), whereby a journal would guarantee to publish replications of studies that they have published (unless there is a demonstrated significant methodological flaw with the initial study). This could incur a large commitment from journals. However, if publishers are no longer bound by printed page limitations, such initiatives become more feasible (e.g., Wiley-Blackwell has removed page limitations for many of its journals). Another step is for more journals to explicitly comment on the acceptability of null findings because one hindrance to replication is that not reproducing the initial statistically significant findings may leave authors vulnerable to negative reviews from the authors of the initial study or from general bias against null findings. One direct way of reducing such bias is via a results-free peer review at Stage 1 (Button, Bal, Clark, & Shipley, 2016), where authors seek reviews on the basis of rationale, methods, and planned analyses alone and, once approved, the full manuscript with results is submitted for a Stage 2 review (e.g., *BMC Psychology*). Although mitigating against bias at review, such a mechanism cannot reduce problems earlier in the research process because the data are already known to the researcher, so questionable research practices (e.g., such as hypothesizing after results are known, *p* hacking) could still have happened prior to the results-free review. Thus, journals that encourage submission of replication studies and carry out a results-free review attain only Level 2 of the TOP Guidelines on replication.

A mechanism that aims to address these problems, as well as to increase the amount and quality of replication, is the article type referred to as Registered Reports (see Marsden et al., 2018). Registered Reports were pioneered by the journal *Cortex* in 2013 and have been adopted by about 66 journals as a permanent article type (https://cos.io/rr) at the time of writing. For Registered Reports, a manuscript receives an initial (Stage 1) review of the study purpose, aims, materials, data collection, and analysis protocols. Crucially, the Stage 1 review occurs before the data are collected. If approved, the materials and procedures are time stamped as a preregistration and given formal in-principle acceptance by the editor (Nosek & Lakens, 2014). Then, data collection, analysis, and report writing proceed and are submitted for a Stage 2 review. At this stage, as the design and methods were approved beforehand, studies cannot be "reviewed out" due to assertions relating to methodological flaws. Thus, in-principle acceptance incentivizes researchers to undertake a replication by reassuring them with a pledge of publication prior to investing in the data collection.

It is unsurprising, therefore, that to date Registered Reports include a high proportion of replication studies (see the list at Center for Open Science, 2017), relative to the proportions found in standard publication routes as observed in the current and previous studies. Indeed, journals that offer Registered Reports as a route to publishing replication research meet the highest level (Level 3) of the TOP Guidelines on replication. Thus, Registered Reports have the potential to address many of the observations in our synthesis, including (a) few replications of studies with null findings, (b) low rate of publication of replications overall, (c) lack of direct replications, (d) extensive and unacknowledged heterogeneity between initial and replication studies, and (e) associations between supportiveness of replications' findings and author overlap or materials availability.

Registered Reports also carry other benefits: (a) they allow peer review to inform the study at the design stage (rather than when it can be too late to improve the study); (b) they reduce questionable research practices; and (c) they accommodate any methods where data collection, coding, and analyses can be predetermined (e.g., including observations and interviews). Registered Reports do not preclude additional exploratory data collection or analysis because authors can report these in addition to the registered protocol and analysis, although such exploratory endeavors would be subject to review at Stage 2. Importantly, there is also potential to adapt the procedure to fit uniquely to exploratory designs and associated epistemologies in Exploratory Reports (McIntosh, 2017).

> Recommendation 12:
> Encourage journal editorial boards to consider accepting Registered Report article types and, where this is not possible, to consider undertaking results-free reviews.

Another barrier to replication is that it is difficult to include both a replication and an extension study within one published article given normal space limitations, yet this study structure may alleviate the stigma attached to doing replications. We found few examples of such article types (e.g., Barcroft & Sommers, 2005; Marsden et al., 2013) and they were not included in our current synthesis because we investigated replications of studies in different publications. Current limitations on article length are probably one reason why this was rare, but we are hopeful that this situation will change as publishers remove formal word limits as publication moves online (though to the best of our knowledge, only Wiley-Blackwell has yet done this). There are at least two

models: One is a study that begins with a direct replication of a study published previously, followed by a partial or conceptual replication (to ascertain generalizability, boundary conditions, etc.), and another is an initial study followed by a confirmatory direct replication to test the robustness of the original data and methods. We recommend both of these routes.

> Recommendation 13:
> Encourage publishers to lift word limits or provide online capacity to encourage more replication work within individual study reports.

Our data demonstrate that the perceived low prestige of replication research is unfounded in at least two respects: perceived ease and perceived low impact. Carrying out well-justified, carefully administered replications that are rigorously analyzed in relation to their initial study is no trivial task and very rare in self-labeled replications to date. Our data also show that replications have been relatively highly cited and have been published in some of the highest impact journals. Further, the three journals that we found to have published the highest number of replications were found by Plonsky et al. (2017) to have the highest perceived prestige. As a community, researchers can further enhance the impact and prestige of replications by co-citing them along with their initial studies (see the proposals by Koole & Lakens, 2012, for incentivizing replication). Beyond enhancing the impact of replications, such a practice would reflect a valid and comprehensive reporting of the state of the literature because readers would know the extent to which the results of the initial study are reliable or generalizable.

> Recommendation 14:
> When the initial study is cited, also cite (at least any direct and partial) replication studies of it.

Wider Cultural Changes in Academia

Changing the incentives in our wider academic culture is even more challenging than changing the editorial, review, and citation practices discussed above. Of course, a driving force to shape behavior is funding (as noted by Baker, 2015, and Collins, 1985). Although we found that replication studies to date have not uniquely been cheap and easy studies (because a reasonable proportion had relatively costly characteristics, such as oral measures, classroom environments, and longitudinal and intervention designs), we found few replications using expensive equipment, corroborating concerns expressed by Laws (2016). Combined with the low rates of replication research overall, this

indicates that funding mechanisms are indeed critical for improving replication effort. However, incentivization in academia tends to be entrenched in rewarding originality (Chambers, 2017). For example, approximately 60% of centrally distributed funding of UK universities is allocated on the basis of the three criteria of originality, rigor, and significance of research (Research Excellence Framework, 2011). Although replication studies could score highly on rigor and significance because these are arguably inherent in good replication work, they are likely to score lower on originality. Nevertheless, changes to the most recent Research Excellence Framework (2017)—for example, a reduced number of published outputs and reward for open science practices—could incentivize large multisite preregistered replication projects. We further note five recent funding initiatives that should help replication effort. Two of these directly promote replication research: the IRIS Replication Award for published replications that used materials from IRIS and the Netherlands' Organization for Scientific Research scheme dedicated to funding replication studies (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, 2017). The other three initiatives indirectly promote replication efforts by adopting a Registered Reports approach to review: *Language Learning*'s Early Career grant scheme, which prioritizes one award for a Registered Report (Marsden et al., 2018); funder collaborations with journals that integrate the Registered Reports model of peer review into the grant funding process, such as The Children's Tumour Foundation with *PLOS ONE* (2017) and the charity funder Cancer Research UK with the journal *Nicotine and Tobacco Research* (Munafò, 2017).

> Recommendation 15:
> Increase funding from institutional through to international levels to promote replication as an integral part of the research process.

Professional associations could also incorporate replication strands into their conference programs, endorse replication as a valued part of tenure applications, and encourage reporting standards and publication practices that facilitate replication (e.g., see American Educational Research Association, 2006, 2011). The American Association for Applied Linguistics (2017) recently amended its guidelines to recommend that "high quality replication studies, which are critical in many domains of scientific inquiry within applied linguistics, be valued on par with non-replication-oriented studies." Engaging students with conducting replication studies has also been discussed (see Frank & Saxe, 2012; Porte, 2012), and we are aware of several graduate programs where replication is an integral part of training and assessment.

Recommendation 16:
Encourage efforts (e.g., via teaching and training infrastructures, institutional recognition, and professional association conferences and promotion guidance) to reward those who include replication research in their work.

Conclusion

We conclude by considering a few key implications for the future metascience and production of replication research. This includes acknowledgements of some of the limitations of our study and arguments. First, we hope this study will stimulate replications and extensions of the systematic review itself. Also, when more direct replications are available, future syntheses will be able to investigate the extent of reproducibility in the field quantitatively. That is, rather than using author interpretations and subjective ratings as was appropriate and necessary in the current study, meta-analytic techniques would be appropriate for examining reproducibility in direct replications (where high reproducibility is clearly expected) and for assessing the effects of any operational heterogeneity (recalling that intentional heterogeneity is sometimes designed and predicted to yield nonreproduced findings).

Second, we do not suggest that increased replication alone will improve the reliability and validity of all L2 research. To some extent, we agree with the argument of Schmidt and Oh (2016) that rather than increasing replication, other issues, such as publication bias and questionable research practices, need to be tackled first, and then meta-analyses could address the lack of direct replication (see Coyne, 2016, for related arguments, and Schimmack, 2016, on the value of replicability indices to detect likely publication bias in lieu of actual replication studies for investigating reproducibility). Although we agree that these other issues require attention, we argue that meta-analysis could not address the lack of replication. As a retrospective mechanism, meta-analysis cannot address some problems that can be addressed by replication, such as lack of parity between studies, which reduces the critical mass of adequately powered comparable studies that answer sufficiently similar questions to be included in any meta-analysis (Laws, 2016).

Third, understanding the causes of low levels of published self-labeled replication requires data about the experiences and opinions of editors, reviewers, and researchers, using questionnaires and interviews. This would reveal the extent to which the observed lack of replication originates in low levels of execution, self-labeling, article submission, and/or actual publication. That is, we do not have a good understanding of the extent to

which replications are in fact submitted to journals but rejected and if so, why.

Finally, our key finding is perhaps the very low number (67) of self-labeled replication studies in L2 research, an especially striking finding when it is set against the 50 calls and commentaries on replication in the field. All four reviewers requested that we express in stronger terms the perturbing limited amount and quality of self-labeled replication research. Using the words of one reviewer, we could sum up our data as

> providing an unequivocal view of the state of replication research in the field: It is disparate, loose, rare, flawed, inconsistent, and opaque. If a foundation of high-quality replication studies is a prerequisite for a healthy discipline, the field of second language research occupies very hazardous terrain.

We have identified many factors that must work together to change production of and attitudes toward replications, including increased transparency of materials and data, multisite collaboration, more consistent self-labeling of replications, fewer and more transparent alterations of features from one study to the next, and increased publication via article types such as Registered Reports. Recommending that these and other practices be incorporated more systematically into our communities is intended to propel us toward a more mature field, whose terrain embraces replication research that is more convergent, tighter, more frequent, less flawed, more consistent, and more transparent.

Final revised version accepted 4 December 2017

Notes

1 We acknowledge that this is not the most satisfactory approach to citing and referencing large multisite collaborations as it does not identify the lead authors in the text and cannot list all the collaborators in the reference, thus removing Mikołajczak, S., Moreno, N., Slabakova, R. as per APA convention.

2 An anonymous reviewer wondered whether we might undertake this, but because measures and other variables were very often changed between the initial and replication studies, and other changes were unacknowledged, quantifying a general level of reproducibility in existing L2 self-labeled replication would not have been informative. This is in contrast to recent endeavors in the field of psychology that have undertaken new, direct replications with the explicit goal of measuring reproducibility.

3 Where a study replicated more than one initial study ($k = 7$), we coded according to the replicators' aims and analyses. If two initial studies were replicated

separately (because each initial study had different aims and designs and because analyses in the replication were presented separately), then these were coded as unique initial–replication pairs (Chen, 2011; Cobb, 2003; Robinson, 2005). On the other hand, if two initial studies were replicated because the initial studies had very similar designs and aims and the replication was presented as if replicating one collapsed study, then this was coded as a single initial–replication pair (e.g., DeKeyser & Sokalski, 1996; Ellis et al., 2014; Liu, 1985; Walters, 2012). Five studies included in a review by Polio (2012b) were not included in our study because they did not self-label clearly as a replication study in a journal article's title or abstract or because they were not a replication of a study reported in a separate publication.

4 This is generous for two main reasons. First, the calculation is based only on journals that have published self-labeled replications rather than on all journals that have ever published L2 research. The list of the latter would be very difficult to estimate, and it would probably provide an unfair representation of replication rate because it would include an extremely wide range of journals across multiple disciplines. The field of L2 research does not have an SSCI discipline-specific list, such as the ones used for psychology and education by Makel et al. (2012) and Makel and Plucker (2014), respectively. Second, our start date is from the earliest replication published rather than the start date of each journal (e.g., *The Modern Language Journal* began publishing in 1916).

5 "We do not discourage contributions that present null results" (*SLR*) and "Lack of statistically significant results, or difficulty in drawing clear conclusions, will not necessarily rule out publication of interesting contributions" (*LTest*).

6 Google Scholar includes citations from many types of publications, including books (unlike the Web of Science used by Makel & Plucker, 2014).

7 The mean length of time elapsed since replications were published was 13.1 years (*mode* = 11). For initial studies, it was 20.5 years (*mode* = 21).

8 Impact did not seem to be affected by whether the replication's findings tended to support the initial study's findings or not: citations *not/partially not supportive*, $M = 6.35$ ($SD = 8.609$, $k = 19$), citations *partially/very supportive*, $M = 7.62$ ($SD = 5.742$, $k = 46$), $U = 330.5$, $z = -1.536$, $p = .124$, $d = 0.190$, 95% CI [–0.3456, 0.7254].

9 Intervention was defined for coding purposes as "an experimental manipulation to cause learning, beyond normal practice."

10 We coded our sample studies for their research areas, and the data are available in Appendix S2 in the Supporting Information online and at http://www.iris-database.org. We found a very wide range of subdomains of research, and the coding was subjective, involving multilayered coding categories. We could not discern any patterns in terms of particular areas that had more or fewer replication studies.

11 Studies that did not use language learners collected data from, for example, teachers, corpora, or textbooks.
12 Where age ranges were given, the median was used.
13 This was calculated by dividing the study sample size by the number of groups (or conditions) in each study. The calculation excluded two pairs of studies that gathered large-scale data from formal tests—these replications increased the sample size of the initial studies by 44,612 and 1,415.
14 We were unable to locate the mean for a direct comparison, though a personal communication indicated this was 35 ($SD = 64$, 95% CI [30, 40]).
15 We acknowledge that these ratings are subjective, but the technique is very similar to, though slightly more fine-grained than, the 3-point scale used by Makel et al. (2012) and Makel and Plucker (2014): success, failure, mixed. This approach was fit for our purpose because, unlike the recent large-scale replication efforts in psychology that set out to statistically assess reproducibility across multiple studies by conducting a direct replication of each of the initial studies, we aimed to provide a review of replications—of all kinds—that had already been conducted.
16 Other studies provided (partial) eta squared on omnibus tests or were coded other/unclear/not applicable. Ellis et al. (2014) used regression coefficient beta, another standardized measure of the magnitude of effect size.
17 In cases where there was no authorship overlap, several replications ($k = 14$) included the initial authors in the acknowledgements (which sometimes indicates academic lineage/collaboration). Combining overlap in authorship with a mention in the acknowledgements yielded almost equal numbers of studies with authorship commonalities ($k = 33$) and those with none ($k = 34$).
18 In terms of association between supportiveness and replications being in the same journal as the initial study, our data did not suggest a strong trend (see Table S3-7 in Appendix S3 in the Supporting Information online). This is broadly in line with Makel and Plucker (2014). However, a dichotomous coding of supportiveness (same journal: 24% not supportive vs. 76% supportive; different journal: 32% not supportive vs. 64% supportive) suggests that this direction of investigation may be worth pursuing once the field has a larger body of direct replications.
19 Likelihood ratio for small samples, $LR(2) = 11.052$, $p = .004$; Fisher's exact test because one cell (16.7%) had cell count of fewer than 5, $p = .005$. Five studies were excluded because cell counts were too small. Only two studies provided all the instruments used to collect all data used in the analysis (one supportive and one not). One study provided all instruments (supportive), and two studies could not be coded as to whether they were supportive or not.

References

American Association for Applied Linguistics. (2017). *Promotion and tenure guidelines*. Retrieved November 1, 2017, from http://www.aaal.org/?page=PT

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33–40.

American Educational Research Association. (2011). Code of ethics. *Educational Researcher*, *40*, 145–156. https://doi.org/10.3102/0013189X11410403

Au, T. K. (1983). Chinese and English counterfactuals: The Sapir-Whorf hypothesis revisited. *Cognition*, *15*, 155–187. https://doi.org/10.1016/0010-0277(83)90038-0

Au, T. K. (1984). Counterfactuals: In reply to Alfred Bloom. *Cognition*, *17*, 289–302. https://doi.org/10.1016/0010-0277(84)90012-X

Author guidelines for contributors. (2017). *The Modern Language Journal*. Retrieved March 1, 2017, from http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1540-4781/homepage/ForAuthors.html

Bakan, D. (1967). *On method*. San Francisco: Jossey-Bass. https://doi.org/10.1177/001316446802800431

Baker, M. (2015). Over half of psychology studies fail reproducibility test. *Nature News and Comment*. Retrieved March 1, 2017, from https://doi.org/10.1038/nature.2015.18248

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. https://doi.org/10.1177/1745691612459060

Barcroft, J., & Sommers, M. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*, 387–414. https://doi.org/10.10170/S0272263105050175

Basturkmen, H. (2014). Replication research in comparative genre analysis in English for academic purposes. *Language Teaching*, *47*, 377–386. https://doi.org/10.1017/S0261444814000081

Berez-Kroeker, A., Gawne, L., Kung, S., Kelly, B., Heston, T., Holton, G., et al. (2017). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, *56*, 1–18. https://https://doi.org/10.1515/ling-2017-0032

Bergmann, C., Meulman, N., Stowe, L. A., Sprenger, S. A., & Schmid, M. S. (2015). Prolonged L2 immersion engenders little change in morphosyntactic processing of bilingual natives. *Neuroreport*, *26*, 1065–1070. https://doi.org/10.1097/WNR.0000000000000469

Bohannon, J. (2014). Replication effort provokes praise and "bullying" charges. *Science*, *344*, 788–789. https://doi.org/10.1126/science.344.6186.788

Branco, A., Cohen, K.-B., Vossen, P., Ide, N., & Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: Introducing an LRE special section. *Language Resources and Evaluation 51*, 1–5. https://doi.org/10.1007/s10579-017-9386-7

Button, K. S., Bal, L., Clark, A., & Shipley, T. (2016). Preventing the ends from justifying the means: Withholding results to address publication bias in peer-review

[Editorial]. *BMC Psychology*, *4*, 59. https://doi.org/10.1186/s40359-016-0167-7

Center for Open Science. (2017). *Registered reports*. Retrieved January 16, 2018, from https://cos.io/rr/

Chambers, C. (2016, November 9). *Accountable replication policy at Royal Society Open Science* [Blog post]. Retrieved November 1, 2017, from http://neurochambers.blogspot.co.uk/2016/11/an-accountable-replication-policy-at.html

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press. https://doi.org/10.1177/1745691613514450

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*, 30–49. https://doi.org/10.1006/jmla.2001.2832

Chen, Y. (2011). Studies on bilingualized dictionaries: The user perspective. *International Journal of Lexicography*, *24*, 161–197.

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review*, *59*, 393–424. https://doi.org/10.3138/cmlr.59.3.393

Collins, H. M. (1985). The possibilities of science policy. *Social Studies of Science*, *15*, 554–558. Retrieved November 1, 2017, from http://www.jstor.org/stable/285370

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology*, *4*, 28. https://doi.org/10.1186/s40359-016-0134-3

DeKeyser, R. M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, *46*, 613–642. https://doi.org/10.1111/j.1467-1770.1996.tb01354.x

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly*, *50*, 132–153. https://doi.org/10.1002/tesq.217

Devlin, H. (2016, September 21). Cut-throat academia leads to "natural selection of bad science," claims study. *Science Guardian*. Retrieved November 1, 2017, from https://www.theguardian.com/science/2016/sep/21/cut-throat-academia-leads-to-natural-selection-of-bad-science-claims-study

Dimroth, C., Rast, R., Starren, M., & Watorek, M. (2013). Methods for studying a new language under controlled input conditions: The VILLA project. *Eurosla Yearbook*, *13*, 109–138. https://doi.org/10.1075/eurosla.13.07dim

Earp, B. D. (2016). What did the OSC replication initiative reveal about the crisis in psychology? An open review of the draft paper entitled "Replication initiatives will not salvage the trustworthiness of psychology" by James C. Coyne. *BMC Psychology*, *4*, 1–19. Retrieved November 1, 2017, from https://www.researchgate.net/publication/293651901_What_did_the_OSC_replication_initiative_reveal_about_the_crisis_in_psychology

Earp, B. D., Everett, J. A., Madva, E. N., & Hamlin, J. K. (2014). Out, damned spot: Can the "Macbeth Effect" be replicated? *Basic and Applied Social Psychology*, *36*, 91–98. https://doi.org/10.1080/01973533.2013.856792

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 1–11. https://doi.org/10.3389/fpsyg.2015.00621

Let's replicate [Editorial]. (2006). *Nature*, *442*, 330. https://doi.org/10.1038/442330b

Ellis, N. C., & Sagarra, N. (2010). The bounds of adult language acquisition. *Studies in Second Language Acquisition*, *32*, 553–580. https://doi.org/10.1017/S0272263110000264

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition*, *33*, 589–624. https://doi.org/10.1017/S0272263111000325

Ellis, N. C., Hafeez, K., Martin, K. I., Chen, L., Boland, J., & Sagarra, N. (2014). An eye-tracking study of learned attention in second language acquisition. *Applied Psycholinguistics*, *35*, 547–579. https://doi.org/10.1017/S0142716412000501

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172. https://doi.org/10.1017/S0272263105050096

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication research's disturbing trend. *Journal of Business Research*, *60*, 411–415. https://doi.org/10.1016/j.jbusres.2006.12.003

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904. https://doi.org/10.1007/s11192-011-0494-7

Faretta-Stutenberg, M., & Morgan-Short, K. (2011). Learning without awareness reconsidered: A replication of Williams (2005). In G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. P. Botana, & E. Rhoades (Eds.), *Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions* (pp. 18–28). Somerville, MA: Cascadilla Proceedings Project.

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLOS ONE*, *10*, e0118053. https://doi.org/10.1371/journal.pone.0118053

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, *27*, 537–554. https://doi.org/10.1177/0265532209354771

Fitzpatrick, T., & Meara, P. (2004). Exploring the validity of a test of productive vocabulary. *Vigo International Journal of Applied Linguistics*, *1*, 55–74.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*, 975–991. https://doi.org/10.3758/s13423-012-0322-y

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*, 595–599. https://doi.org/10.1177/1745691612460686

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PLOS ONE*, *12*, e0175583. https://doi.org/10.1371/journal.pone.0175583

Han, C. (2016). Reporting practices of rater reliability in interpreting research: A mixed-methods review of 14 journals (2004–2014). *Journal of Research Design and Statistics in Linguistics and Communication Science*, *3*, 49–75. https://doi.org/10.1558/jrds.29622

Hartshorne, J., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience*, *6*, 1–14. https://doi.org/10.3389/fncom.2012.00008

Hubbard, R., & Armstrong, J. S. (1994). Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing*, *11*, 233–248. https://doi.org/10.1016/0167-8116(94)90003-5

Instructions for contributors. (2017). *Studies in Second Language Acquisition*. Retrieved March 1, 2017, from https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/information/instructions-contributors

Ioannidis, J. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again . . . *Science*, *334*, 1225–1225. https://doi.org/10.1126/science.334.6060.1225

Kahneman, D. (2014). A new etiquette for replication. *Social Psychology*, *45*, 310.

Kanno, K. (2000). Case and the ECP revisited: Reply to Kellerman and Yoshioka (1999). *Second Language Research*, *16*, 267–280. https://doi.org/10.1191/026765800672956803

Kelly, C. W., Chase, L. J., & Tucker, R. K. (1979). Replication in experimental communication research: An analysis. *Human Communication Research*, *5*, 338–342. https://doi.org/10.1111/j.1468-2958.1979.tb00646.x

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., et al. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*, e1002456. https://doi.org/10.1371/journal.pbio.1002456

Kim, J., & Nam, H. (2017). Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition*, *39*, 431–457. https://doi.org10.1017/S0272263115000510

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives in Psychological Science*, *7*, 608–614. https://doi.org/10.1177/1745691612462586

Larson-Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *The Modern Language Journal*, *101*, 244–270. https://doi.org/10.1111/modl.12386

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*, 127–159. https://doi.org/10.1111/lang.12115

Laws, K. R. (2016). Psychology, replication and beyond [Editorial]. *BMC Psychology*, *4*, 30. https://doi.org/10.1186/s40359-016-0135-2

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345–366. https://doi.org/10.1093/applin/amu040

Lee, S.-K., & Huang, H.-T. (2008). Visual input enhancement and grammar learning. *Studies in Second Language Acquisition*, *30*, 307–331. https://doi.org/10.10170/S0272263108080479

Lindsay, S. (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832. https://doi.org/10.1177/0956797615616374

Lindsay, S. (2017). Sharing data and materials in psychological science [Editorial]. *Psychological Science*, *28*, 699–702. https://doi.org/10.1177/0956797617704015

Lindstromberg, S. (2016). Inferential statistics in language teaching research: A review and ways forward. *Language Teaching Research*, *20*, 741–768. https://doi.org/10.1177/1362168816649979

Liu, L. (1985). Reasoning counterfactually in Chinese: Are there any obstacles? *Cognition*, *21*, 239–270. https://doi.org/10.1016/0010-0277(85)90026-5

Luijendijk, H., & Koolman, X. (2012). The incentive to publish negative studies: How beta-blockers and depression got stuck in the publication cycle. *Journal of Clinical Epidemiology*, *65*, 488–492. https://doi.org/10.1016/j.jclinepi.2011.06.022

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159.

MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*, *67*, 254–275. https://doi.org/10.1111/lang.12220

Makel, M., & Plucker, J. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304–316. https://doi.org/10.3102/0013189X14545513

Makel, M., Plucker, J., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science*, *7*, 537–542. https://doi.org/10.1177/1745691612460688

Markee, N. (2017). Are replication studies possible in qualitative second/foreign language classroom research? A call for comparative re-production research. *Language Teaching*, *50*, 367–383. https://doi.org/10.1017/S0261444815000099

Marsden, E. (in press). Open science and methodological transparency in applied linguistics research. In C. Chapelle (Ed.), *Encyclopaedia of applied linguistics*. Oxford, UK: Blackwell.

Marsden, E., & Mackey, A. (2014). IRIS: A new resource for second language research. *Linguistic Approaches to Bilingualism*, *4*, 125–130. https://doi.org/10.1075/lab.4.1.05mar

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York: Routledge. https://doi.org/10.4324/9780203489666

Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. (2018). Introducing Registered Reports at *Language Learning*: Promoting transparency, replication, and a synthetic ethic in the language sciences [Editoral]. *Language Learning*, *68*.

Marsden, E., Thompson, S., & Plonsky, L. (in press). A methodological synthesis of self-paced reading tests in second language research. *Applied Psycholinguistics*.

Marsden, E., Williams, J., & Liu, X. (2013). Learning novel morphology: The role of meaning and orientation of attention at initial exposure. *Studies in Second Language Acquisition*, *35*, 619–654. https://doi.org/10.1017/S0272263113000296

Marsman, M., Schönbrodt, F., Morey, R., Yao, Y., Gelman, A., & Wagenmakers, E. (2017). A Bayesian bird's eye view of "Replications of important results in social psychology." *Royal Society Open Science*, *4*, 160426. https://doi.org/10.1098/rsos.160426

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*, 523. https://doi.org/10.3389/fpsyg.2017.00523

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. https://doi.org/10.1037/a0039400

McIntosh, R. (2017). Exploratory reports: A new article type for Cortex. *Cortex*, *96*, A1–A4. https://doi.org/10.1016/j.cortex.2017.07.014

McManus, K., & Marsden, E. (2017). Online and offline effects of L1 practice in L2 grammar learning: A partial replication. *Studies in Second Language Acquisition*. Published online June 19, 2017. https://doi.org/10.1017/S0272263117000171

Mellers, B. A., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275. https://doi.org/10.1111/1467-9280.00350

Meulman, N., Wieling, M., Sprenger, S. A., Stowe, L., & Schmid, M. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. *PLOS ONE*, *10*(12), e0143328. https://doi.org/10.1371/journal.pone.0143328

Morgan-Short, K., Heil, J., Botero-Moriaty, A., & Ebert, S. (2012). Allocation of attention to second language form and meaning: Revisiting the use of think aloud

protocols. *Studies in Second Language Acquisition*, *34*, 659–685. https://doi.org/10.1017/S027226311200037X

Morgan-Short, K., Marsden, E., Heil, J., Issa, B., Leow, R. P., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R., & Szudarski, P. (2018). Multi-site replication in SLA research: Attention to form during listening and reading comprehension in L2 Spanish. *Language Learning*, *68*.

Munafò, M. (2017). Improving the efficiency of grant and journal peer review: Registered Reports funding. *Nicotine & Tobacco Research*, *19*, 773. https://doi.org/10.1093/ntr/ntx081

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*. https://doi.org/10.1038/s41562-016-0021

Nakamura, D. (2012). Input skewedness, consistency, and order of frequent verbs in frequency-driven second language construction learning: A replication and extension of Casenhiser and Goldberg (2005) to adult second language acquisition. *International Review of Applied Linguistics in Language Teaching*, *50*, 1–37. https://doi.org/10.1515/iral-2012-0001

National Academies of Sciences, Engineering, and Medicine. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*. Washington, DC: National Academies Press. https://doi.org/10.17226/21915

Neulip, J. W., & Crandall, R. (1993). Everyone was wrong: There are lots of replications out there. *Journal of Social Behavior and Personality*, *8*, 1–8.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning*, *50*, 417–528. https://doi.org/10.1111/0023-8333.00136

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, *65*, 470–476. https://doi.org/10.1111/lang.12104

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, D., Breckler, S. J., et al. (2015a). *TOP guidelines*. Retrieved November 1, 2017, from https://cos.io/top

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015b). Promoting an open research culture. *Science*, *348*, 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B., & Lakens, D. (2013). Call for proposals special issue of social psychology on "Replications of important results in social psychology." *Social Psychology*, *44*, 59–60. https://doi.org/10.1027/1864-9335/a000143

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *43*, 137–141. https://doi.org/10.1027/1864-9335/a000192

National Science Foundation. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science: Report of the subcommittee on*

replicability in science. Advisory committee to The National Science Foundation directorate for social, behavioral, and economic sciences. Retrieved November 1, 2017, from http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

Nederlandse Organisatie voor Wetenschappelijk Onderzoek. (2017). *Replication studies*. Retrieved November 1, 2017, from https://www.nwo.nl/en/research-and-results/programmes/replication+studies

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. https://doi.org/10.1126/science.aac4716

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning 63*(s1), 1–24. https://doi.org/10.1111/j.1467-9922.2012.00735.x

Oswald, F., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. https://doi.org/10.1017/S0267190510000115

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*, 539–544. https://doi.org/10.1177/1745691616646366

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, *61*, 993–1038. https://doi.org/10.1111/j.1467-9922.2011.00663.x

Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 116–132). New York: Cambridge University Press.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L. (2015). Statistical power, p values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23–45). New York: Routledge.

Plonsky, L., Blair, R., Boyce, K., Kim, A., Li, F., Qi, D., et al. (2017). *Quality, prestige, and impact in L2 research journals*. Manuscript in preparation.

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*, 267–278. https://doi.org/10.1177/0267658314536436

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*, 538–553. https://doi.org/10.1111/modl.12335

Plonsky, L., Egbert, J., & LaFlair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, *36*, 591–610. https://doi.org/10.1093/applin/amu001

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

PLOS. (2015). Positively negative: A new PLOS One collection focusing on negative, null and inconclusive results [Web log article]. *The Missing Pieces*. Retrieved November 1, 2017, from http://blogs.plos.org/collections/collections/the-missing-pieces

PLOS ONE. (2017). The Children's Tumour Foundation and PLOS ONE announce a new funder-publisher partnership [Press release]. Retrieved November 1, 2017, from http://www.ctf.org/news/ctf-plos-one-funder-publisher-partnership

Polio, C. (2012a). No paradigm wars please! *Journal of Second Language Writing*, *21*, 294–295. https://doi.org/10.1016/j.jslw.2012.05.008

Polio, C. (2012b). Replication in published applied linguistics research: A historical perspective. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 47–91). New York: Cambridge University Press.

Polio, C., & Gass, S. (1997). Replication and reporting: A commentary. *Studies in Second Language Acquisition*, *19*, 499–508.

Porte, G. (2012). *Replication research in applied linguistics*. New York: Cambridge University Press.

Porte, G., & Richards, K. (2012). Focus article: Replication in second language writing research. *Journal of Second Language Writing*, *21*, 284–293. https://doi.org/10.1016/j.jslw.2012.05.002

Research Excellence Framework. (2011). *Assessment framework and guidance on submissions*. Retrieved March 1, 2017, from http://www.ref.ac.uk/pubs/2011-02

Research Excellence Framework. (2017). *Initial decisions on REF 2021*. Retrieved November 1, 2017, from http://www.hefce.ac.uk/pubs/year/2017/CL,332017

Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfeld, and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, *27*, 235–268.

Rohrer, D., Pashler, H., & Harris, C. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General*, *144*, e73–e85. https://doi.org/10.1037/xge0000058

Rosenthal, R. (1979). The "file drawer" problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Santos, T. (1989). Replication in applied linguistics research. *TESOL Quarterly*, *23*, 699–702. https://doi.org/10.2307/3587548

Schimmack, U. (2016, January 31). *The replicability-index: Quantifying statistical research integrity* [Blog post]. Retrieved March 1, 2017, from https://replicationindex.wordpress.com/2016/01/31/a-revised-introduction-to-the-r-index

Schmid, M. S. (2011). *Language attrition*. Cambridge, UK: Cambridge University Press.

Schmid, M., Berends, S. M., Bergmann, C., Brouwer, S. M., Meulman, N., Seton, B. J., et al. (2015). *Designing research on bilingual development: Behavioral and neurolinguistic experiments*. London: Springer.

Schmidt, F., & Oh, I. S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, *4*, 32–37. https://doi.org/10.1037/arc0000029

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100. https://doi.org/10.1037/a0015108

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. https://doi.org/10.1016/j.jesp.2015.10.001

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U. (2016). Each reader decides if a replication counts: Reply to Schwarz and Clore (2016). *Psychological Science*, *27*, 1410–1412. https://doi.org/10.1177/0956797616665220

Smith, B., & Lafford, B. A. (2009). The evaluation of scholarly activity in computer-assisted language learning. *The Modern Language Journal*, *93*, 868–883. https://doi.org/10.1111/j.1540-4781.2009.00978.x

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice-versa [Editorial]. *American Statistician*, *49*, 108–112. https://doi.org/10.1080/00031305.1995.10476125

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71. https://doi.org/10.1177/1745691613514450

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 435–452). New York: Russell Sage Foundation.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–336. https://doi.org/10.1111/j.1467-1770.1994.tb01104.x

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279–301). Amsterdam: John Benjamins.

Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology. *Theory & Psychology*, *26*, 540–548. https://doi.org/10.1177/0959354316637136

Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and Type 1 error: The domain problem. *New Ideas in Psychology*, *45*, 19–27. https://doi.org/10.1016/j.newideapsych.2017.01.002

Trenkic, D., Mirkovic, J., & Altmann, G. (2014). Real-time grammar processing by native and non-native speakers: Constructions unique to the second language. *Bilingualism: Language and Cognition*, *17*, 237–257. https://doi.org/10.1017/S1366728913000321

Trofimovich, P., & Ellis, N. (2015). Open science badges [Editorial]. *Language Learning*, *65*, v–vi. https://doi.org/10.1111/lang.12134

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110. https://doi.org/10.1037/h0031322

Vandergrift, L., & Cross, J. (2017). Replication research in L2 listening comprehension: A conceptual replication of Graham & Macaro (2008) and an approximate replication of Vandergrift & Tafaghodtari (2010) and Brett (1997). *Language Teaching*, *50*, 80–89. https://doi.org/10.1017/S026144481500004X

VanPatten, B. (2002a). Processing Instruction: An update. *Language Learning*, *52*, 755–803. https://doi.org/10.1111/1467-9922.00203

VanPatten, B. (2002b). Processing the content of input-processing and processing instruction research: A response to DeKeyser, Salaberry, Robinson, and Harrington. *Language Learning*, *52*, 825–831. https://doi.org/10.1111/1467-9922.00205

VanPatten, B., & Cadierno, T. (1993a). Input processing and second language acquisition: A role for instruction. *The Modern Language Journal*, *77*, 45–57. https://doi.org/10.1111/j.1540-4781.1993.tb01944.x

VanPatten, B., & Cadierno, T. (1993b). Explicit instruction and input processing. *Studies in Second Language Acquisition*, *15*, 225–243. https://doi.org/10.1017/S0272263100011979

VanPatten, B., & Oikkenon, S. (1996). Explanation versus structured input in processing instruction. *Studies in Second Language Acquisition*, *18*, 495–510. https://doi.org/10.1017/S0272263100015394

VanPatten, B., & Williams, J. (2002). *Research criteria for tenure in second language acquisition: Results from a survey of the field*. Unpublished manuscript, University of Illinois at Chicago.

Walters, J. (2012). Aspects of validity of a test of productive vocabulary: Lex30. *Language Assessment Quarterly*, *9*, 172–185. https://doi.org/10.1080/15434303.2011.625579

Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. *System*, *25*, 261–274. https://doi.org/10.1016/S0346-251X(97)00013-4

Wicherts, J., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS One*, *6*, e26828. https://doi.org/10.1371/journal.pone.0026828

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Wong, W. (2001). Modality and attention to meaning and form in the input. *Studies in Second Language Acquisition*, *23*, 345–368. https://doi.org/10.1017/S0272263101003023

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Included Studies, Plus Commentaries and Exclusion Criteria.
Appendix S2. Data Coding Spreadsheet with Data.
Appendix S3. Additional Tables Summarizing Results.